UNIVERSIDAD AUTÓNOMA DE MADRID

TESIS DOCTORAL

Collective Intelligence in Estimation Problems: How Groups Achieve It and How to Build Techniques to Improve It

---

Autor:
Gabriel Madirolas Pérez

Director:
Dr. Gonzalo García
de Polavieja Embid

Memoria de Tesis doctoral presentada
para la obtención del título de
Doctor
por la Universidad Autónoma de Madrid
Programa de Doctorado en Biofísica
Departamento de Física de la Materia Condensada
Facultad de Ciencias

Junio de 2017

Dedicado a Olga, la luz de mi vida
Dedicated to Olga, the light of my life

Table of contents

List of figures

List of tables

Acknowledgements

When I tried to figure out a list of people to acknowledge for my thesis, I started to suspect that I've had no merit in its completion. The truth is that there is going to be more people out of this words than in. And it is not going to be oversights. I feel that every single person with which I have shared a word or a gesture has helped me construct a story that converges into this volume.

But there is a handful of people which has not only been of help, but have consciously inverted part of all their energy in making the path worth walking. The path that started some five years ago in Madrid when I talked to my teacher Fernando Sols about my will to enter academia. He put me in touch with Gonzalo García de Polavieja, who glimpsed something in me that even me was unable to glimpse. He gave me the opportunity to join his lab despite my lack of experience and high qualifications.

In Gonzalo's Lab at the Cajal Institute I met some of the most amazing people I could have met. Julián, Robert, Alfonso, Sara, Ángel and Raúl made me feel like if I was going to job in the living room of my home. Latter Marta and Paco joined and added the freshness of youth to the road. And Ivonne joined our lives to stay.

Then Lisboa came, and a full city of sea and light seeped through my senses. And life never was never the same again. In Lisbon I found, besides one million wonderful people, a whole bounce of enthusiastic learners and livers. How easy is to spend time abroad when there you have people like Victòria, Tomas, Andres, Ivar, Antonia, Mattia, Fran and Cris.

And how about all the people in the past that in one ways or anothers, piece by piece, experience by experience, have contributed to make be who and how I am. Ignacio who is past, present and certainly future. My Robóticos Jordi and Jandro. Miguel, Dani and every single musician I have played with. All friends I have made (and enemies if that's the case) in El Escorial, Mallorca and Menorca.

Thanks to all the people I have met in my life, because from every single one I have obtained the opportunity to learn what to do and not to do, what I want and what I don't want to be my life. And when I have not learned, it is because I did not make the effort!

But it is the ones that are always there who you rely more on. My brother Gustavo, blood of my blood but above all mind of my mind. His wonderful wife Josefina and two diamonds Miranda and Diana. To my late grandmas Carmen and Sofía. And to their two sons, who are my parents. They taught me the passion for life, for love, for knowledge, for music, for endless learning.

Not everyone is lucky to find a second family. But I am certain that no one has ever been so lucky to find the family I have found. They are there in dark and light hours. Tania, Roque and Rodrigo have been brothers to me, and their families a joy in my travels to this miraculous land called Burgos. Olga's parents, Maribel and Ángel, have been an example of unconditional help, support, understanding, acceptance of one's convictions and love.

Olga is all. Without her there is no life, there is no love. I would not be half of the half the man I am. Thanks to her I am writing these words. Not a single day since we are together I have not felt her passion, her art, her tenderness, her endless determination. She took me out of the well and pushed me to the top of a mountain. I could not correspond in a hundred life, but she deserves trying with all heart and soul.correspond in a hundred life, but she deserves trying with all heart and soul.

Summary

In this thesis we have addressed the subject of collective intelligence in humans, in particular the aggregation of individual estimates to improve over the average of the opinions of the group.

In the first chapter we modeled the distribution of opinions in an estimation task, and how it changes upon the knowledge of the estimates of other subjects. To do that we adapt a previous model of a two choices decision task in animals, based on Bayesian estimation and probability matching, to the case of humans estimating a quantity that can take a semi-continuous range of values. The model predicts a weighted geometric mean for the aggregation of private and social information.

Once the model for a collective is proposed, we translate it to the point of view of an individual reconsidering his first estimate when receiving the opinions of the others. We take the advantage of this adaptation to investigate the presence of individuality on the social behavior of each subject. Doing that we obtain a criterion to classify subjects according to their resistance to social influence.

We investigate whether there exists a correlation between behavioral externalization of confidence in own opinion, and degree of expertise of the subject in the question estimated. To do that we use two methods. A first method was computing the average of a significantly big enough subgroup of the subjects that resist more to social influence. The second method is based on finding subgroups in the joint distribution of the numerical estimates and the measure of resistance to social information. We found that for four out of six questions, the method provides a better estimate to the true value than the average of the estimates of the whole group.

Finally we compared the performance of the selection rule based on the behavioral confidence with the selection of the subjects that declare higher confidence in their answer on a 1 to 6 scale. We find that there is no correlation between the two measures of confidence,

and that declared confidence provides no significant improvement over the average of the opinions of the group.

In the second chapter we investigated the possibility of applying the predictions of our model to the case of groups of subjects reaching a consensus estimate after a free discussion period. To do that we analyzed data from an experiment in which subjects were first asked to estimate the number of sweets in a jar, and then arranged in groups of three to discuss and give a common guess about the same jar.

We tried to find models of a weighted geometric mean of pre-discussion estimates, and deduced it gives infinite solutions for the case of three subjects. Then we tried to fit models in which the same weight is applied to the two closer pre-discussion estimates, or in which the distance between the latter acts as a measure of the attraction over the further one (spring model). We also tried a costs model which quantifies how costly is for each subject to move from his original opinion to the consensus one.

The last of the mentioned models suggested that we should study a diversity in strategies used by the groups to reach the consensus estimates. We calculated the likelihood of the consensus decision being reached by a noisy use of each of eight simple aggregation strategies, and found that the geometric mean is the more likely strategy to be generating the experimental results.

We finally analyzed whether there was an influence of the pre-discussion estimates configuration and range in the preference for a certain strategy. We found that when there was a big asymmetry in pre-discussion estimates, with one being much higher than the others, the geometric mean was clearly dominant. In those situations, the geometric mean was the strategy that provided the more accurate prediction of the actual number of sweets in the jar, therefore showing that groups detected it as the more beneficial strategy and predominantly applied it.

In the third chapter we investigate the possibility of applying new techniques of Machine Learning, particularly neural networks, to experimental data consisting of independent judgments of doctors over possible cases of skin cancer. The network was only fed with the diagnosis and declared confidences of the doctors, and not the images with which they diagnosed, therefore the network only was used to aggregate collective information.

We trained the network over half of the experimental cases, and validated its performance over the other half. We found that the network provides better accuracy than heuristics

already known to provide beneficial aggregation of estimates. The network even outperforms aggregation rules that already combine the accuracy and confidence of each doctor.

The results were tested in groups of two, three, five and seven doctors, and we found that the network consistently outperforms any of the aggregation heuristics for any group size. We even found that the network increases its performance with group size faster than any of the other strategies, except the majority voting heuristic.

Resumen

En esta tesis nos hemos aproximado al tema de la inteligencia colectiva en humanos, en particular a la agregación de estimaciones individuales para mejorar sobre el promedio de las opiniones del grupo.

En el primer capítulo modelizamos la distribución de las opiniones en una tarea de estimación, y como cambia al conocer las estimaciones de los otros. Para ello adaptamos un modelo anterior para experimentos de decisión forzada entre dos o varias opciones, basado en la estimación Bayesiana y el 'probability matching', al caso de humanos estimando una cantidad que pueda tomar un rango de valores cuasi-continuo. El modelo predice una media geométrica ponderada para la agregación de información privada y social.

Una vez que el modelo para un colectivo es propuesto, lo trasladamos al punto de vista de un individuo reconsiderando su primera estimación al recibir las opiniones de los otros. Aprovechamos esta adaptación para investigar la presencia de individualidad en el comportamiento social de cada sujeto. Mediante ello obtenemos un criterio para clasificar a los sujetos de acuerdo a su resistencia a la influencia social.

Investigamos si existe una correlación entre exteriorización conductual de la confianza en la opinión propia, y el grado de precisión del sujeto en la cuestión estimada. Para eso utilizamos dos métodos. Un primer método fue calcular el promedio de un subgrupo de un tamaño suficientemente significativo de los sujetos que resisten más a la influencia social. El segundo método está basado en encontrar subgrupos en la distribución conjunta de las estimaciones numéricas y la medida de resistencia a la información social. Encontramos que para cuatro de seis cuestiones, los métodos proporcionan una mejor estimación de valor correcto que el promedio de las estimaciones del grupo.

Finalmente comparamos la precisión de la regla de selección basada en la confianza conductual con la selección de los sujetos que declaran una mayor confianza en su respuesta en una escala de 1 a 6. Encontramos que no hay una correlación entre las dos medidas de

confianza, y que la confianza declarada no proporciona una mejora significativa sobre el promedio de las opiniones del grupo.

En el segundo capítulo investigamos la posibilidad de aplicar las predicciones de nuestro modelo al caso de grupos alcanzando una estimación de consenso tras un periodo de discusión libre. Para ello analizamos datos de un experimento en el que los sujetos fueron primero invitados a estimar el número de dulces en un bote, y entonces reunidos en grupos de tres para discutir y dar una estimación común sobre el mismo bote.

Tratamos de encontrar modelos de una media geométrica de las estimaciones previas a la discusión y deducimos que existe un número infinito de soluciones para el caso de tres sujetos. Entonces ensayamos modelos en los que se aplica el mismo peso a las dos estimaciones más cercanas de las previas, o en el que la distancia entre estas actúa como una medida de atracción sobre la estimación más lejana (modelo de muelle). También tratamos de aplicar un modelo de costes que cuantifica cuan costoso es para cada sujeto moverse desde su opinión inicial a la de consenso.

El último de los mencionados métodos sugirió que deberíamos estudiar la diversidad en estrategias utilizadas por los grupos para alcanzar las decisiones de consenso. Calculamos la plausibilidad de que la decisión de consenso sea alcanzada por un uso ruidoso de cada una de ocho estrategias de agregación simples, y encontramos que la media geométrica es la que tiene una mayor probabilidad de estar generando los resultados experimentales.

Finalmente analizamos si existe una influencia de la configuración de las estimaciones previas a la discusión y su rango por la preferencia en alguna estrategia determinada. Encontramos que cuando hay una asimetría en la estimaciones previas, con una mucho más alta que las otras, la media geométrica es claramente dominante. En esa situación, la media geométrica era la estrategia que proporcionaba una predicción más precisa del número real de dulces en el bote, mostrando por lo tanto que los grupos la detectan como la estrategia más beneficiosa y la aplican.

En el tercer capítulo investigamos la posibilidad de aplicar las nuevas técnicas de 'Machine Learning', particularmente redes neuronales, a datos experimentales consistentes en los juicios de médicos sobre posibles casos de cáncer de piel. La red solo recibía como datos de entrada los diagnósticos y la confianza declarada por los médicos, y no las imágenes con las que se había diagnosticado, y por lo tanto solo era utilizada para agregar la información colectiva.

Entrenamos la red con la mitad de los casos experimentales, y validamos su precisión con la otra mitad de los casos. Encontramos que la red proporciona una mejor predicción que heuríistcas de las que ya se sabía que proporcionaban una agregación de la información beneficiosa. La red incluso mejora a reglas de agregación que ya combinan la precisión y la confianza de loa médicos.

Los resultados fueron comprobados en grupos de dos, tres, cinco y siete médicos, y encontramos que la red consistentemente supera a cualquiera de las heurísticas de agregación para cualquier tamaño de grupo. Encontramos incluso que la red mejora su precisión con el tamaño de grupo de una manera más rápida que las otras estrategias, excepto la de tomar la opinión mayoritaria.

Chapter 1

General Introduction

Absolute certainty about the state of the world is a desirable but in practice unreachable goal. In the absence of information about the environment, a thinking being has to rely on its previously stored knowledge. Conversely, as the individual is confronted with a new challenge that demands a new approach, or at least involves some unknown variables, new information must be gathered from the outside world. One way to obtain this new information is exploration with the senses and motion. However, social species have the ability to take advantage of the knowledge already possessed by other conspecifics. This knowledge is easily transmitted by the use of strategies specific for the communication between members of the same species, which have been selected and refined during the evolutive process.

Collectives in general, and human collectives in particular, have been extensively shown to outperform individuals on wide variety of tasks and everyday problems, including migrations, foraging, unknown quantity estimation and market forecasting. It is in the spirit of polls and general elections. Particularly, the method of averaging the individual estimates of the members of the group, or the majority voting have been proven beneficial in solving many complex issues. However, there are many situations where different drawbacks can appear that make the naive aggregation tasks less than optimal. In this situations, it will be desirable to find methods for extracting a better knowledge from the collective (estimates), or at least to prove that such a thing is possible.

There are many reasons that can lead a collective to a biased average opinion. Sometimes the distribution of knowledge across the population is not homogeneous, and cluster around wrong values are created. This clusters of mistaken individuals may be caused by the existence of different approaches to solving a question, to the influence of mass media, or even to urban legends.

Sometimes the problem is not just that the average is biased, which in practice will always happen at lower or higher degree, but that the cost of making mistakes is so that it is better to choose the opinion of one subject at random than the average of the group. When costs function, which expresses the relationship between the mistakes and how much is lost due to the distance to the true value, are not convex, Jensen's inequality don't hold. In that cases, averaging the opinions of the group does not guarantee to reduce the cost over selecting one subject at random.

One evident method of for improving the group estimate is to detect those individuals that posses a higher lever of expertise in the field under evaluation. This could be obtained via a questionnaire, or having access to historical performance on tasks as much similar to the current as possible. However, in many situations the problem solvers that face a question are nearly novel to it. Other times it is the problem itself that can not be found a similar in the past. For that situations, it is desirable to have methods only based on the present behavior of the subject, which might include his answers to some questionnaires more or less related to the task.

Other method that has been extensively used either in research as in candidate selection is the reported confidence. Although this method has the advantage of not needing access to previous data, it has the drawback of being subjective. That means that sometimes an individual might have a high sense of confidence when providing an answer for a problem for the first time, but see his confidence reduced when faced with details about the issue or with the opinions of other about it.

An increasingly intercorrelated world introduces at the same time a new problem and a solution. On the one hand, it makes the spreading of false beliefs and opinion bias in general easier and faster. On the other hand, it makes available a huge amount of information and opinions from subjects with a high diversity of backgrounds and points of view. Dealing with that big amount of information requires advanced techniques of computation and aggregation. The advantage of methods that take into account correlations between the estimators that are overlooked by traditional aggregation heuristics is that only the useful pieces of information are taken into account.

In this work we try to draw a path that starts with the simplest aggregation rules like averaging or majority voting, goes through more complex rules that take into account the distribution of knowledge between the group, and ends up posing new techniques that reflect and take advantage of complexities and subtleties unreachable with traditional approaches.

Chapter 2

Modeling Social Influence and Improving Collective Estimations

2.1 Introduction

In this chapter we will show two first approaches to the main subjects of this thesis. First, we introduce the mathematical modeling of the aggregation of social information. To do that we will present a model of how subjects within a collective integrate information about other's beliefs with their own previous private knowledge. Second, we will propose methods to improve the estimation of the collective taking advantage of the individuality among the group of estimating subjects.

In the study of collective intelligence in humans the work of Francis Galton is deservedly considered a milestone. In 1907, the anthropologist and statistician was the first to experimentally demonstrate the advantages of collective estimation (Galton [7]). At a farmers' fair, he found that the median of the independent estimates made by 784 farmers of the weight of a slaughtered ox was better than any of their individual estimates. Since then, collective estimations, computed as mean, median or geometric mean values of the group, have been shown to improve upon the estimates of most individuals of a group in several different contexts, an effect popularly known as *Wisdom of Crowds* (hereinafter WOC) (Surowiecki [30], Page [23], Lee and Shi [15], Wagner and Vinaimont [32], Easley and Kleinberg [6], Krause et al. [12], King et al. [9]).

However, human crowds can also be notoriously bad at making collective estimations for many estimation tasks (Krause et al. [12]). For example, in tasks requiring memory or mental

calculation, subjects often give estimates with large deviations from factual values (Lorenz et al. [18]). Of course, methods to overcome a biased crowd average have been proposed (Whitehill et al. [33], Mannes et al. [20], Budescu and Chen [5], Zhou et al. [35])

Social interactions can have an additional negative effect in biased crowds (King et al. [9], Lorenz et al. [18]). When individuals learn the estimations of the other members of the group, they typically change their own estimation towards the more common values. After social influence, the collective has thus a distribution of estimations more strongly peaked around the biased solution. This can give the collective perception of an agreement but the value agreed upon can be far from the truth (Lorenz et al. [18]).

We propose to turn the negative effect of social interactions to our advantage and improve collective estimations. We do so by taking into account the individuality of the members of the group. Francis Galton argued for each individual counting the same in the collective estimation (Galton [7]). But for situations in which most individuals are strongly biased, we would be in a better position with methods selecting the unbiased individuals. Of course, this can be done by finding how well each individual performs in a domain of knowledge and weight them accordingly for similar tasks (Wolfers and Zitzewitz [34], Whitehill et al. [33], Lee et al. [16]). Here we do not consider the case of access to a classification of individuals by performance. Instead we used the impact of social interactions on estimations to extract individuals in the following way. We first obtained a model of estimation in a collective and used it to measure how much each individual of the collective resists social influence.

To model a human estimation task, we considered subjects as estimators of some quantity according to the information available to them. We wanted to model the integration of social information with the previous information that the person has and was the background for his initial estimation. This modeling approach had already been successfully used to study social behavior in fish and ants that chose among a low number of options (Pérez-Escudero and de Polavieja [26], Arganda et al. [1]). We adapted this low number of choices model to the case in which subjects had to estimate an unknown quantity.

we then reasoned that not all individuals should be influenced equally by the public information. We used our model to classify each individual by their resistance to social influence as a measure of confidence on their private information. We then used these values of social resistance obtained from the model to extract the subgroup of people resisting social influence, and found that they give an improved collective estimation. Our proposal is then

to use the geometric mean of the estimations of individuals with high social resistance as a better estimator than the WOC, as we here show for the dataset from reference (Lorenz et al. [18]).

2.2 Results

2.2.1 Model for a collective

When a group is asked to estimate individually a quantity that can take in principle any positive value, it has been found that the distribution of independent estimates (previous, for example, to any social information) is a log normal (Lorenz et al. [18]). We thus decided to take the logarithm of the raw estimates x_i emitted by the subjects, because the new variable $y_i \equiv \log x_i$ is then normally distributed and is therefore easier to manipulate analytically. The Gaussian that fits the estimates will be characterized by a mean μ_p and a standard deviation σ_p, denoting $y \sim N(\mu_p, \sigma_p)$, where the subscript p stands for "private" and refers to the fact that the estimation has been made only based on the private information that each subject already had. Our model predicted a distribution of estimates emitted by the subjects after social interactions of the form (see Section 2.4.2 for deduction of the expression)

$$f_Y(y) = N\left(w_p\mu_p + w_s\mu_s, \sqrt{1-w_s}\sigma_p\right). \tag{2.1}$$

That is, the predicted distribution f_Y in Eq 2.1 is still a Gaussian $y \sim N\left(\mu_f, \sigma_f\right)$ on the logarithm of estimates, but the parameters have changed according to the integration of social with private information. The mean μ_f is a combination of the mean μ_p of the private distribution and a parameter μ_s that encapsulates numerically the social information that subjects have received. The relative influence of private and social information in the final mean is expressed in Eq 2.1 in the form $\mu_f = w_p\mu_p + w_s\mu_s$, with the strength of each factor given by private and social weights, w_p and w_s, with values between 0 and 1 and with $w_p + w_s = 1$.

Parameter μ_s in Eq 2.1 can take many functional forms, that will depend on the specific content of the social information and the particular circumstances under which it is provided to the subject. Regarding only social information consisting on estimates made by other participants, the data provided can consist for example in one or various estimates emitted by previous participants, or in the best estimate so far, or the arithmetic or geometric mean

of the previous estimates Lorenz et al. [18], King et al. [9]. Each of these cases could in principle enter Eq 2.1 not only with a different numerical value μ_s, but also with a different social weight w_s depending on how strongly does this particular case of information affect the subject's opinion.

We considered principally two types of social information. The first is based on the situation where a report of the estimates of the members of the group is provided. In that case we derived that the form of μ_s will be $\mu_s \equiv \log x_s$, with x_s the geometric mean of the n estimates provided to the subject (see Sections 2.4.2 and 2.4.3 for deduction of the expression):

$$x_s = \left(\prod_{i=1}^{n} x_i \right)^{1/n}. \tag{2.2}$$

The other form of interaction we analyzed arises when the subject is provided only with the arithmetic mean of the previous estimates. In that case, the social parameter μ_s will be computed directly extracting the logarithm of the arithmetic mean (see also Section 2.4.3 for deduction):

$$x_s = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{2.3}$$

Logically, in cases where not the detailed estimates nor the arithmetic mean, but the geometric mean of previous estimates is provided to the subject, the form of μ_s would be also given by Eq 2.2, but the social weight in 2.1 might be different in both cases.

The impact of the two mentioned social interactions is different in the mean of the final distribution of estimates, but equal in the standard deviation. In the case given by Eq 2.2, as the expected value of the geometric mean (x_s in our case) of a sample extracted from a distribution that follows a log-normal is the median ($\exp(\mu_p)$) of the population, we will have on average that $x_s = \exp(\mu_p)$. Therefore, the mean of the distribution of estimates is expected not to change due to this particular form of social interaction. On the other hand, when the subjects are provided with the arithmetic mean of previous estimates, which expected value is $x_s = \exp\left(\mu_p + \sigma_p^2/2\right)$, the final distribution is expected to have parameter $\mu_f = \mu_p + w_s\sigma_p^2/2$. That is, as the mean of a log-normal distribution is higher than the median, the final distribution is shifted towards higher values than the original distribution previous to social influence. In the two information cases we considered, the standard deviation is predicted to be reduced in a proportion given by $\sigma_f = \sigma_p\sqrt{1-w_s}$, with a higher reduction the higher is the social weight w_s. That means that not only social influence can

sometimes change the average opinion of the group, but is expected always to make the group be in higher agreement than before social interactions.

2.2.2 Test of the collective model

We tested the distribution predicted in 2.1 with data collected in a group experiment which is described in more detail in section 2.4.1 (and was presented for the first time in Lorenz et al. [18]). Concerning the test of our model, it is enough to say here that 12 groups of 12 subjects were asked six questions about geographical and social facts. For example, subjects were asked to estimate the length of the border between Switzerland and Italy. Each member of the group made a first independent and anonymous guess of the actual value being asked about. Then information about the guesses of the other members of the groups was given to each individual, under two information conditions. The "full information" condition consisted in showing the subject, after he had made his first estimation, a report with the twelve estimates made by the members of his group. The "aggregated information" condition consisted in providing after the first estimation only the arithmetic mean of the estimates made by the group. Of the six questions proposed, two were made under the "full information" condition and two under the "aggregated information". The other two questions were made under a "no information" condition, where no information about the estimations of the individuals was provided at any moment, that served as a control condition. Then, each individual was asked to reconsider individually his answer to the question, and to emit a new estimate (that could be the same as his first).

The number of estimates per question, information condition and iteration was 24, corresponding to two experimental groups of 12 subjects. To gain statistical power, we decided to pool together all the estimates made under each information condition, resulting in 288 estimation processes. But as for each question different means and standard deviations were obtained, we decided to standardize the estimates to the mean and standard deviation of each group and question. Even for a same question and information condition, due to noise the distribution of estimates and therefore the social information produced from it were different from group to group. Formally, if y was the logarithm of the estimate, and μ_p and σ_p was the mean and standard deviation of the estimates in the group (the ones emitted prior to social information), the standardized variable z could be defined by the z-score.

$$z \equiv (y - \mu_p) / \sigma_p \tag{2.4}$$

The distribution of a z-score has by definition zero mean and unitary standard deviation, and so will have the standardized logarithms of each question and group. But we went an step further and expected that, if for every question and group the estimates prior to social information can be considered like extractions from log-normal distributions, when all the standardized distributions are pooled together the resulting distribution should still be distributed according to a standard normal, $z_1 \sim N(0,1)$. The subscript in z_1 indicates that we refer to the first estimation, done prior to knowledge of social information.

For the "full information" condition, after social information the logarithm z_2 of the second estimates standardized according to Eq 2.4 and pooled together are expected to follow a standard normal, transforming according to Eq 2.1 (see Section 2.4.4) to $z_2 \sim N\left(0, \sqrt{1-w_s}\right)$. Statistical analysis on the experimental data confirm that in the "full information" condition the distribution of z-score values before social information cannot be distinguished from a standard normal distribution ($p = 0.36$; Kolmogorov-Smirnov test, Fig 2.1A, blue). All the results of significance tests performed over distributions in this chapter are summarized in Table 2.1. In agreement with the prediction of our model, after social information the standardized variable in the "full information" condition follows a normal distribution ($p = 0.34$; Kolmogorov-Smirnov test, Fig 2.1A, red), with the same mean than before social information ($p = 0.14$; permutations test) but lower standard deviation ($p < 10^{-9}$; permutations test). To obtain the two latter p-values we used a permutations method, based on pooling together the components of two samples and performing random partitions in two groups. This method is explained in more detail in Section 2.4.5, and will be used in the following to obtain p-values unless otherwise stated. Given the analytical form predicted for the final distribution, $N(0, \sqrt{1-w_s})$, from the standard deviation of the data it can be extracted a value of the social weight $w_s = 0.53$.

In the "aggregated information" condition, the distribution of z-score values before social information is also indistinguishable from a standard normal distribution ($p = 0.95$; Kolmogorov-Smirnov test, Fig 2.1B, blue, and Fig 2.2B, blue). The model predicts for the standardized variable after social information not only a narrower distribution, but also a shift in the mean towards higher values (see Section 2.4.4), to $z_2 \sim N(w_s\sigma_p/2, \sqrt{1-w_S})$. However, as the expected value of the final mean depends on the specific standard deviation of the estimates of the initial distributions, we should not in principle pool together the 24 experiments (two questions per each of the twelve groups). Therefore, we first represented separately each of the 24 experiments, and used the σ_p parameter of each one extracted

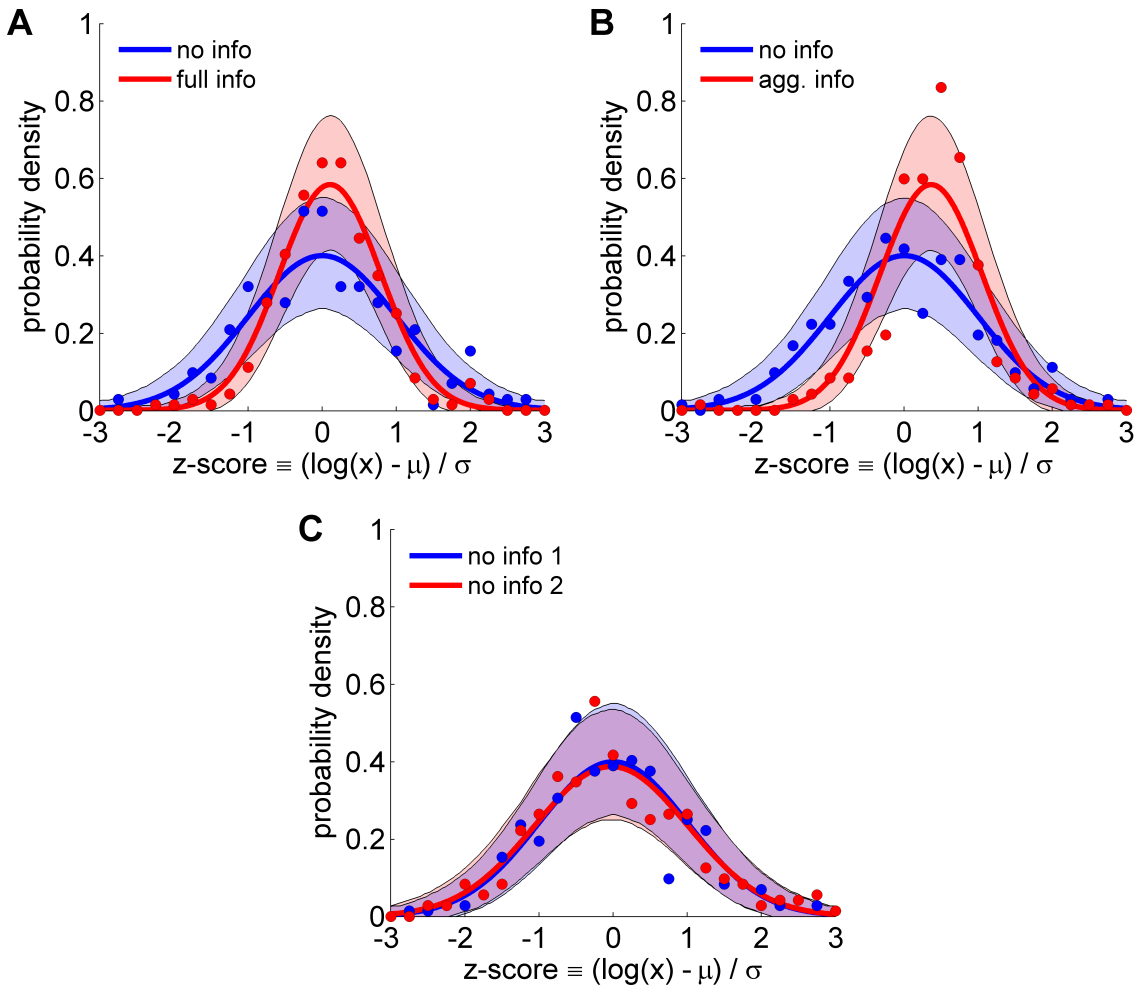

Fig. 2.1 **Comparison of statistical predictions against experiment before and after social information.** (**A**) Probability distribution of estimates before (no info, blue) and after (full info, red) receiving the estimations made by other members of the group. Points are experimental frequencies sampled at intervals of width 0.25 and solid line is a Gaussian fit. Shadowed surface is the area in which 95 per cent of the experiments are expected by the Gaussian fit. The statistical prediction is that after social interactions the distribution of answers is also a Gaussian in the logarithmic domain with the same mean but smaller standard deviation. (**B**) Same as (A) but before (no info, blue) and after (aggregated info, red) giving subjects the mean of the estimates of all subjects. Shadowed surface is the area in which 95 per cent of the experiments are expected using $\overline{\sigma_p} = 1.37$ computed from the 24 experiments, and $w_s = 0.53$ from the "full information" condition. The statistical prediction is that after social interactions the distribution of answers is also a Gaussian in the logarithmic domain with higher mean and smaller standard deviation. (**C**) Same as (A) but estimating twice without interactions in between (first: blue, second: red). The statistical prediction is that after social interactions the distribution of answers is also a Gaussian in the logarithmic domain with the same mean and standard deviation. Data taken from Lorenz et al. [18].

Property	Samples (information condition, trial)	Null Hypothesis Significance Tests	Bayesian tests (95% HDI, accept)
		Kolmogorov-Smirnov	
Normality	no info, t1	0.41, yes	(0.896, 2.04), yes
	no info, t2	0.28, yes	(0.841, 2.01), yes
	full info, t1	0.36, yes	(0.871, 2.05), yes
	full info, t2	0.34, yes	(0.568, 1.51), yes
	agg. info, t1	0.95, yes	(1.07, 2.11), yes
	agg. info, t2	0.052, yes	(0.384, 0.828), no
		Permutations	
Equality of means	no info, t1 - no info, t2	0.75, yes	(-0.138, 0.196), yes
	full info, t1 - full info, t2	0.14, yes	(-0.256, 0.0205), yes
	agg. info, t1 - agg. info, t2	$< 10^{-7}$, no	(-0.551, -0.281), no
Equality of variances	no info, t1 - no info, t2	0.60, yes	(-0.151, 0.0905), yes
	full info, t1 - full info, t2	$< 10^{-9}$, no	(0.203, 0.411), no
	agg. info, t1 - agg. info, t2	$< 10^{-11}$, no	(0.271, 0.475), no
	full info, t2 - agg. info, t2	0.45, yes	(-0.139, 0.0164), yes

Table 2.1 **Kolmogorov-Smirnov, Permutations and Bayesian Significance Tests. Kolmogorov-Smirnov** tests were run with Matlab to check normality. **Permutations** method was performed as explained in Section 2.4.5 to test for the equality of means and equality of variances. For the no difference of means, two sample t-tests were run with Matlab to check compatibility with permutations method. For the no difference of variances, two sample F-tests were run with Matlab with the same purpose. No discrepancies in the acceptance/rejection of the null hypothesis were found in any of the no difference tests. **Bayesian tests** are based on the likelihood of the experimental data given a certain value of the parameters (Kruschke [13]). The method generates a probability distribution of the most credible values of the parameters (or their difference for two distribution comparison) is generated. If a value falls outside the 95% highest density interval (HDI) then it not considered to be a credible value of the parameter or difference of parameters. For the distribution to be considered credibly normal, a value for the degrees of freedom parameter of $\log_{10}(\nu) > \log_{10}(30) \approx 1.48$ is required. Only one discrepancy was found with the null hypothesis methods, and the Bayesian test cannot accept the normality of the estimation distribution generated in the second trial of the "aggregated information" condition. Although the Kolmogorov-Smirnov test did not reject the normality hypothesis, the p-value was slightly above 0.05. In section 2.2.2 and in Fig 2.2 this poor value is explained by the fact that the distribution is better explained by the sum of 24 Gaussians with very similar parameters.

from the set of estimates before social information, and the average w_s parameter of the group extracted from its two "full information" condition questions, to predict the shift in mean and reduction of standard deviation of the distribution after social information in the "aggregated information" condition (Fig 2.2A). The average of the 24 predicted Gaussian distributions predicts correctly the distribution of all the 24 experiments pooled together (Fig 2.2B, red). However, a simpler analysis can be done using an average $\overline{\sigma_p} = 1.37$ computed from the 24 experiments, and $w_s = 0.53$ from the "full information" condition, to predict a distribution $N(w_s\overline{\sigma_p}/2,\sqrt{1-w_s}) = N(0.363, 0.685)$, close to the actual mean and standard deviation of the standardized data (0.393 and 0.655, Fig 2.1B, red). Thus, we found a confirmation of the prediction that when providing the arithmetic mean there will be a shift of the mean of the z-score ($p < 10^{-6}$; permutations test) and a reduction of the standard deviation ($p < 10^{-6}$; permutations test). The model also predicts that , although the means of the distributions after social information might not be equal in the "full information" and the "aggregated information" conditions, the standard deviation of both should be the same ($p = 0.48$, permutations test).

Due to the different expected value of the final standard deviation on each of the 24 experiments in the "aggregated information" condition, the distribution resulting from pooling together the z-score made over the 24 set of estimates need not follow a normal distribution. However, if the σ_p from each experiment are not expected to be very different from each other, the final σ_f of each of the 24 distributions might not be very diverse. This could happen if all the experiments are expected to have similar estimate distributions, all log-normals possibly with different μ_p parameters but similar σ_p. Actually, the z-scores after social information were found to follow a normal distribution, although in the limit of significance ($p = 0.052$, Kolmogorov-Smirnov test). We decided to check all the statistical decisions taken over the distributions with an alternative test. We performed Bayesian tests (Kruschke [13]), that compute how likely are two quantities to take the same value, instead of deciding when they are not found to be different, as done by the significance tests. The results of the tests are shown in Table 2.1. The only decision in which the Bayesian test disagreed with the significance test was precisely in the normality of the final distribution of the z-scores in the "aggregated information" condition, which was found unlikely to be a normal distribution by the Bayesian test. As we have already mentioned, this is due to the distribution being composed of 24 slightly different normal distributions.

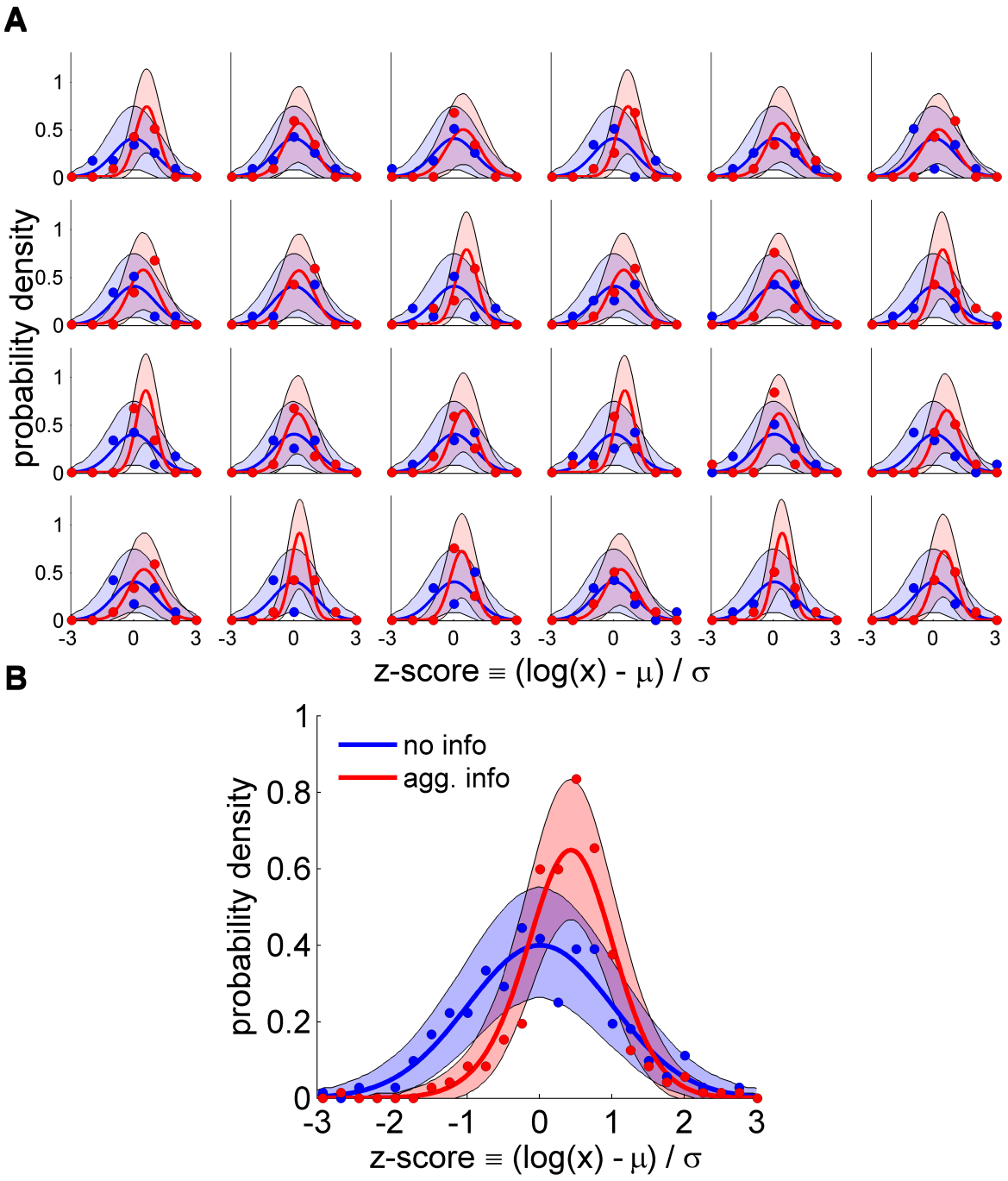

Fig. 2.2 **Distribution of estimates before and after receiving the mean estimation for each experiment.** Same analysis as in Fig 2.1B, but for each of the 24 experiments (**A**) and the sum of the 24 Gaussians (**B**) before (blue) and after (red) receiving the mean value of the estimates. Points are experimental frequencies at intervals of width 1 (A) and 0.25 (B). Shadowed surface is the area in which 95 per cent of the experiments are expected using $\overline{\sigma_p}$ of the group before social interactions and w_s of the group from the "full information" condition. Data taken from Lorenz et al. [18].

In the "no information" condition, subjects repeated the estimation without receiving any information about the estimates of the others (Fig 2.1C). We found normal distributions in both the estimates before social information ($p = 0.41$, Kolmogorov-Smirnov test) and after social information ($p = 0.28$, Kolmogorov-Smirnov test). Moreover, both distributions of z-scores were found to have the same mean ($p = 0.75$, permutations test) and standard deviation ($p = 0.62$, permutations test). This suggests that the changes in the distributions in the "full information" and "aggregated information" conditions were due to the social interactions and not to a repetition of the estimation.

2.2.3 Model for an individual

The model we have presented in the previous sections predicts correctly the behavior of a deciding collective. We studied the most straightforward model of a subject estimating and changing his original opinion that is compatible with the change in the collective estimates distribution shown in Eq 2.1. If an individual emits a first estimate x_1, and then receives social information, his second estimate x_2 can be related with the first via

$$y_2 = w_p y_1 + w_s \mu_s, \tag{2.5}$$

with $\{y_{1,2} \equiv \log x_{1,2}\}$ and μ_s the logarithm of the social information as defined in Section 2.2.1, particularly in the forms expressed in Eq 2.2 or Eq 2.3. This implies that we can make a rough prediction of the second estimate of a subject if we know his first estimate and we have obtained a w_s value for the collective to which the subjects belongs. We found that using the value $w_s = 0.53$ obtained in Section 2.2.2 we can make a good prediction of $\log x_2$ using $\log x_2 = w_p \log x_1 + w_s \mu_s$ (Fig 2.3A). Many previous studies have used the more common linear combination rule $x_2 = w_p x_1 + w_s x_s$ to model the integration of social with prior information that drives the subject to make an estimating decision. The rule we present in Eq 2.5 is a linear combination but in the logarithmic domain, which corresponds when doing the exponentiation on both the left and the right side to a weighted geometric mean between the private opinion and the social information:

$$x_2 = x_1^{w_p} x_s^{w_s}. \tag{2.6}$$

The model in Eq 2.5 or Eq 2.6 is derived from Eq 2.1, which assumes that all subjects in the collective process private and social information using the same weights across the

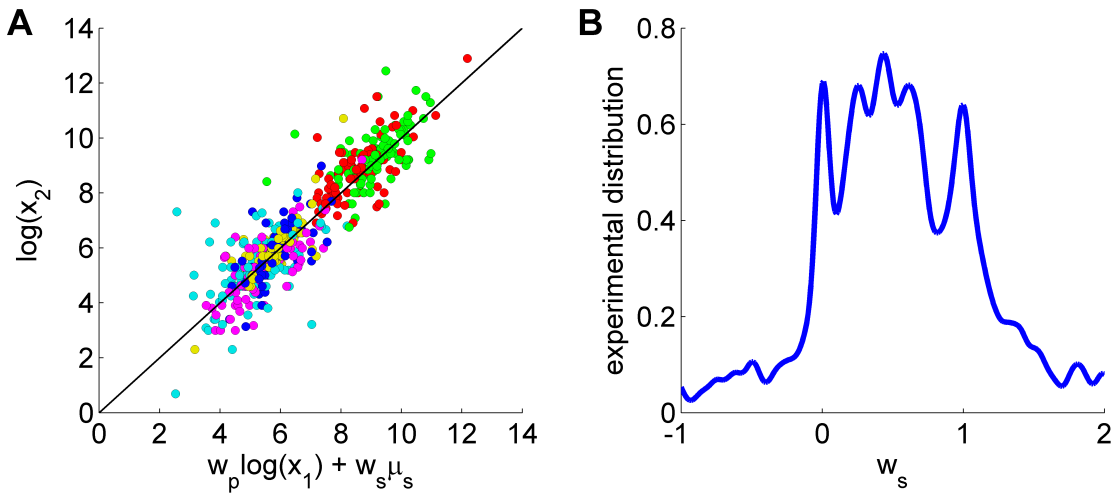

Fig. 2.3 **Average and distribution of social weights.** (**A**) Real vs predicted estimations after social interactions from Eq 2.5 as $\log x_2 = w_p \log x_1 + w_s \log x_s$ using $w_s = 0.53$. Different colors correspond to the six estimation tasks. (**B**) Distribution of experimental social weights with Gaussian kernel smoothing (see Section 2.4.6). Data taken from Lorenz et al. [18].

population. But in the weighted geometric mean of Eq 2.6 there is room to introduce variability in the values of w_p and w_s, as it is an expression for a single individual. We used that by definition $w_p + w_s = 1$ and introduced it into Eq 2.5 to find the value of the social weight that each individual is applying to the social information he receives:

$$w_s = \frac{y_2 - y_1}{\mu_s - y_1}. \tag{2.7}$$

This expression has a very intuitive interpretation. It measures the magnitude of the change in opinion of the subject with a metric given by the distance of his first opinion to the social information he has received. That is, assuming that all his change in opinion can be attributed to an attraction effect by the social information, w_s expresses how much more closer is the subjects opinion to the collective opinion than it was before being informed about it.

To investigate the differences in social weighting across all individuals and questions, we represented the probability distribution across all possible w_s values (Fig 2.3B). We found a striking structure with some noteworthy features. There is a pronounced peak at $w_s = 0$, comprised by all the subjects that resist social influence and do not change their opinion after knowing the other's ($y_2 = y_1$ in Eq 2.7).There is another clear peak at $w_s = 1$, comprised by all the subjects that adopt as their new opinion the value that has been provided to them as social information ($y_2 = \mu_s$ in Eq 2.7). We find a majority of individuals that weight nearly

equally the social information than their own, in correspondence with the estimated value of $w_s = 0.53$ that we found in Section 2.2.2. Finally, there are subjects for which this simple model is not suitable, since they change their opinion to values even further than the social information ($w_s > 1$) or they distance more from it ($w_s < 0$).

2.2.4 Joint density of social weight and estimation

Once we found that there was strong signs of individuality within the distribution of social weights shown in Fig 2.3B, we investigated whether there was a relationship between the social weight applied in Eq 2.6 and the estimates emitted. To do that, we tested the joint density of social weights w_s and estimates $y = \log x_1$ using a Gaussian smoothing of the data (Silverman [28]):

$$f(w_s, y) = \frac{1}{2\pi\sigma_{w_s}\sigma_y n}\sum_{i=1}^{n}\exp\left(-\frac{(w_s - w_{s,i})^2}{2\sigma_{w_s}^2} - \frac{(y - y_i)^2}{2\sigma_y^2}\right), \tag{2.8}$$

with $w_{s,i}$ and $y_i = \log x_{1,i}$ the social weight and private estimate of individual i, respectively, $\sigma_y \equiv \hat{\sigma}_y n^{-1/\gamma_y}$ and $\sigma_{w_s} \equiv \hat{\sigma}_{w_s} n^{-1/\gamma_{ws}}$ being $\hat{\sigma}_y$ and $\hat{\sigma}_{w_s}$ the sample standard deviation of each variable. We used a Gaussian kernel with diagonal covariance, because in principle we did not assume any correlation between social weight and estimation, and because that allowed us to vary the bandwidth applied to one of the variables without altering the other. To find whether the structure detected in Fig 2.3B could translate into a tendency of individuals with different social weights to give different estimations, we varied the resolution coefficient γ_{w_s} while keeping γ_y at an optimal value of $\gamma_y = 6$ (Silverman [28]).

Of the six experimental questions that subjects were asked to estimate, the one concerning the length of the border between Switzerland and Italy showed the most striking behavior under resolution changes in Eq 2.8 (Fig 2.4A). At the lower resolution considered, there is a clear tendency of individuals to give a higher estimate (Fig 2.4A, $\gamma_{w_s} = 6$). As the resolution is reduced, the density splits into two peaks, one at high and low values of w_s (Fig 2.4A, $\gamma_{w_s} = 4, 3$). Finally, at the minimum resolution shown, the two peaks start to blur and the structure breaks (Fig 2.4A, $\gamma_{w_s} = 2$). It is thus clear that for this experimental question, subjects with low w_s tend to give a higher estimate than those with high w_s.

In the question about the number of rapes in Switzerland in 2006, although it can be clearly detected visually a similar trend as in the previous question (Fig 2.4B, $\gamma_{w_s} = 6, 4, 3$), the density ends up breaking without having shown a clear tendency to form two peaks (Fig

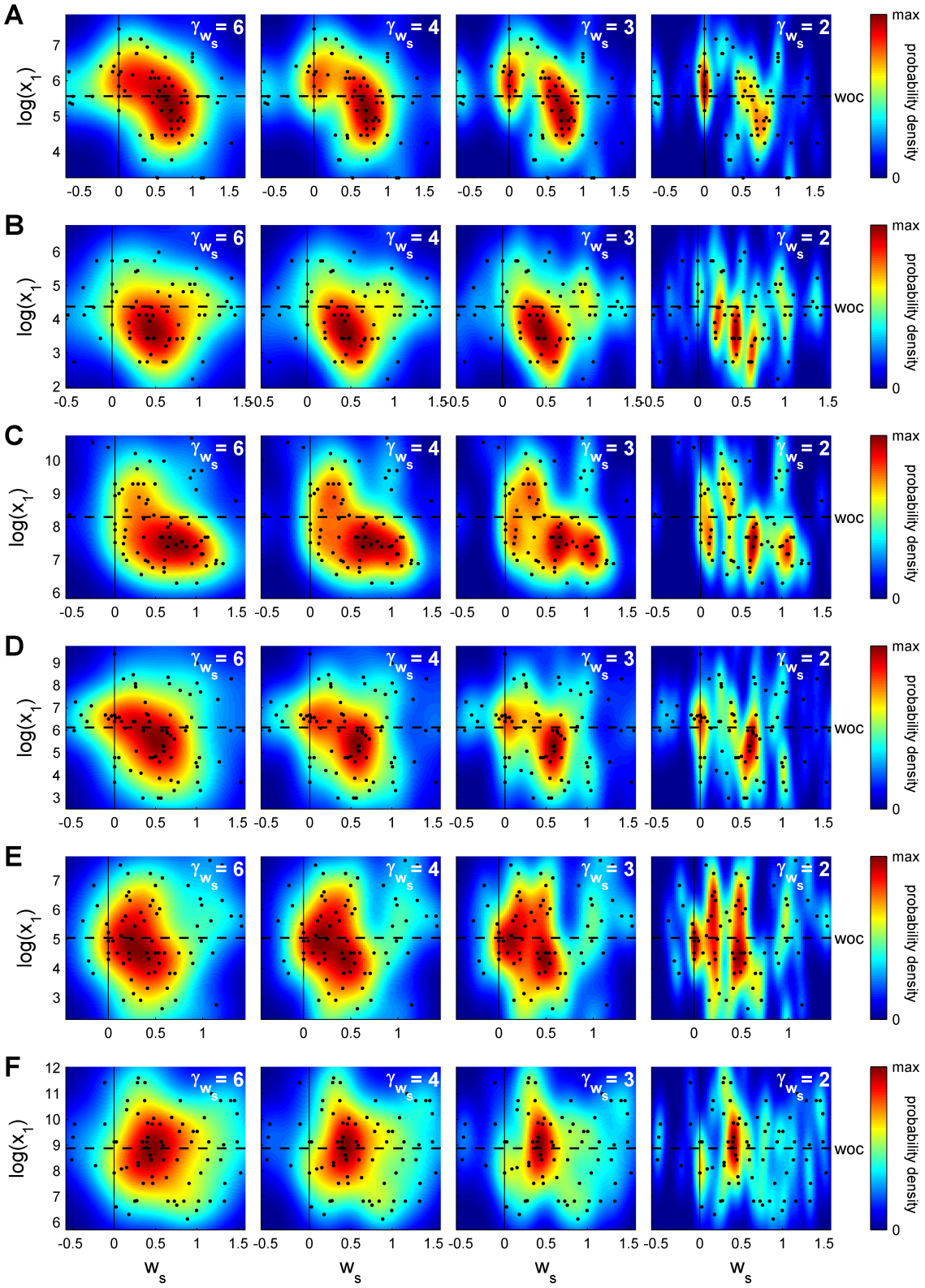

Fig. 2.4 **Joint probability density of social weights w_s and estimates $y = \log x_1$.** The densities are computed with Gaussian smoothing (Eq 2.8) of data (one black dot per individual). Smoothing from lowest resolution in the direction of the social weight w_s ($\gamma_{w_s} = 6$, left) to highest resolution ($\gamma_{w_s} = 2$, right). Question wordings and labels as in Section 2.4.1. Data taken from Lorenz et al. [18].

2.4B, $\gamma_{w_s} = 2$). We will show latter that this difference in the clustering behavior with the previous question will translate into the impossibility of the density of this one to be used for further analysis. Nevertheless, we will still be able to take advantage of the trend to high estimates as w_s decreases.

The questions about the number of assaults in Switzerland in 2006 and the population density of Switzerland show a similar trend as the two previous ones at lower resolution (Fig 2.4C, D, $\gamma_{w_s} = 6$). In both cases is noticeable a tendency to give higher estimates for individuals with low social weight, and this trend is more apparent as the resolution is increased (Fig 2.4C, D, $\gamma_{w_s} = 4,3$). Like in the question about the border length (Fig 2.4A), the population seems to split into a principal group with lower w_s, that gives higher estimates, and another principal group with higher w_s, that gives lower estimates.

In the question about the number of murders registered in Switzerland in 2006 (Fig 2.4E) there is never a clear trend to different estimates at different social weights. At medium resolution (Fig 2.4E, $\gamma_{w_s} = 3$) there are some signs of the formation of two groups, but they almost overlap in the range of $\log(x_1)$ estimates. Moreover, this feature is more apparent at higher resolution (Fig 2.4E, $\gamma_{w_s} = 2$), with the density breaking into more subgroups in addition. In the question about the inhabitant gain in Zurich in 2006 (Fig 2.4F), there is no tendency to different estimates at different social weight values, nor there are signs of the density splitting into two principal groups.

2.2.5 Geometric mean method

We decided to analyze the four questions for which we found in Section 2.2.4 a tendency to give different estimations in subjects with low or high social weight (Fig 2.4A,B,C,D). We hipothesized that one of the two subgroups could be consistently closer to the true answer of the question under estimation.

We then extracted the individuals with lowest social weight. A simple method consists in extracting all individuals with a social weight below the value that gives a result significantly different to the wisdom of the crowd (WOC) value (Fig 2.5A). Specifically, we started from the complete group and its geometric mean as the WOC value. For this case, the WOC value is 302 km (Fig 2.5A). We then eliminated individuals one by one from highest to lowest values of the social weight keeping those with $|w_s| \leq \omega$, with ω a decreasing positive real number. With the remaining individuals, we computed the geometric mean. For ω in the interval between 0.1 and 0.5 of individuals with high resistance to social influence, the

geometric mean increases to values close to 800 km. At the lowest values of ω there is a drop in the geometric mean, but the number of individuals is also low. To isolate the relevant individuals, we found which values of ω give a geometric mean significantly different from the WOC (Fig 2.5A, green dots for $p < 0.05$ and red dots for $p < 0.01$). The significant values of ω are in the interval from 0.06 to 0.45, which correspond to groups whose geometric mean lies between 816 and 464 km, respectively. We then tested that we obtain similar estimations using the complete interval of significant values of ω or only the value of ω giving the highest significance. Specifically, for the complete interval of significant ω we used the following measure that weighted more the values of ω with higher significance as

$$\text{resist 1} \equiv \frac{\int_0^{0.5} q(\omega)\, x_1^{geom}(|w_s| \leq \omega)\, d\omega}{\int_0^{0.5} q(\omega)\, d\omega} \tag{2.9}$$

with $x_1^{geom}(|w_s| \leq \omega)$ the geometric mean of the estimations of individuals with a social weight $|w_s| \leq \omega$, $q(\omega) = 0.05 - p(\omega)$ if the p-value obeys $p(\omega) < 0.05$ and $q(\omega) = 0$ otherwise, and only counting those groups with sufficiently low social weight, $\omega \leq 0.5$. The prediction obtained in this way is 714 km, that deviates only -2.7% from the true value of 734 km while the WOC value of 302 km deviates -59% (Fig 2.5A, 'resist 1', 'truth' and 'WOC'). An alternative to Eq 2.9 would also use the values of ω giving significance but weighted all of them equally, which provides a value of 689 km, -6.2% off the true value (Fig 2.5A, 'resist 2'). Another variant would only take into account a single value of ω with the highest significance ($p = 0.0002$ in this question) that corresponds to $\omega = 0.25$. This gives the prediction of 780 km, 6.3% off the true value (Fig 2.5A, 'resist 3'). The three variants give very similar predictions and a large improvement over the WOC value.

2.2.6 Peaks in the joint distribution method

We also used a second class of methods based on the finding that resisting individuals can form peaks in the joint distribution of estimations and social weight (Fig 2.4A). Methods using the peaks will in general use less individuals but should be valuable when the peaks are clear in the distribution, that is, when they are sharp and separated from other peaks. Specifically, we used clustering by Gaussian mixtures (McLachlan and Peel [22]). The advantage of this method is that, although it depends on the distribution and therefore on the value of the resolution γ_{w_s}, it showed to be very robust to changes in its value. For the question about the length of the Swiss/Italian border, we obtained that the geometric

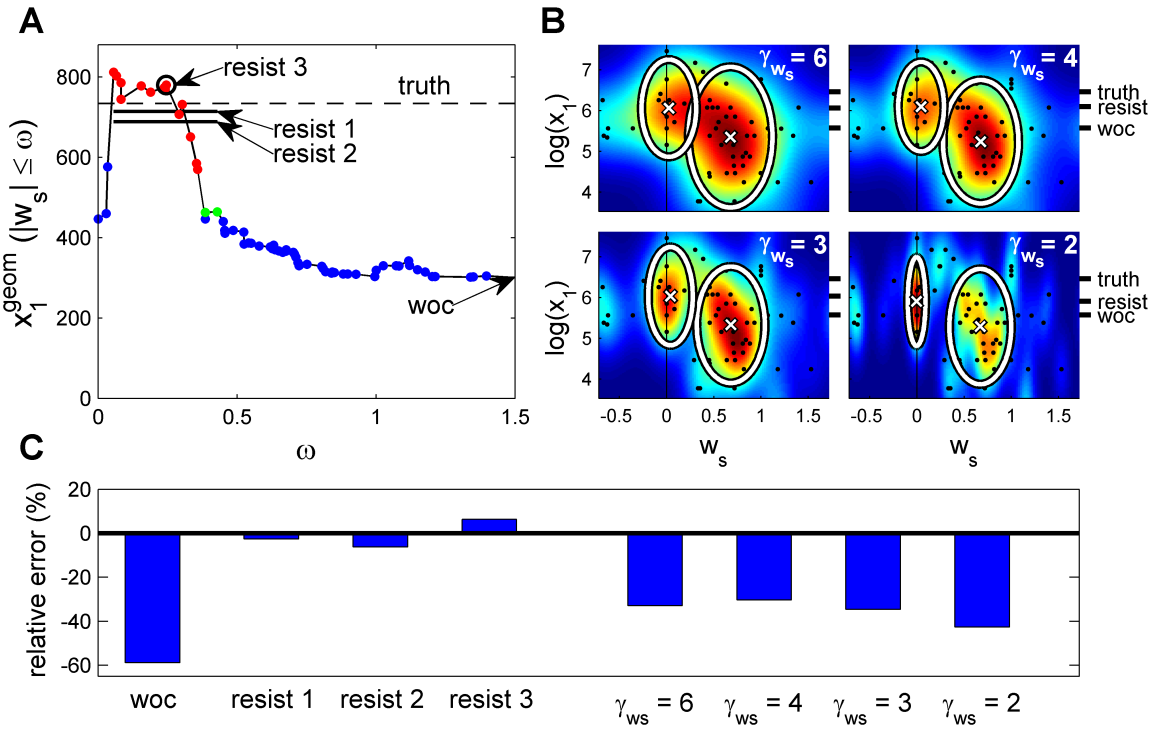

Fig. 2.5 **Wisdom of those resisting social influence for the question "What is length of the Swiss/Italian border?"** (**A**) Geometric mean of estimations for groups containing individuals with social weight $|w_s| \leq \omega$. At low ω the groups are formed by individuals resisting social influence. Blue dots: Groups with prediction not significantly different to wisdom of the crowd (WOC). Green dots: groups significantly different from WOC at $p < 0.05$. Red dots: $p < 0.01$. Value labeled "resist 1" computed from individuals with low social weights and contributing more the values of ω with higher significance (Eq 2.9). Value labeled "resist 2" computed as "resist 1" but not weighting the different ω differently depending on significance levels. Line labeled "resist 3" corresponds to the value of with highest significance. (**B**) Two clusters in the space of estimations and social weights obtained using Gaussian mixtures (McLachlan and Peel [22]). White ellipses delimit the area that contains 95% of the probability density for each of the bivariate Gaussians (Ribeiro [27]). (**C**) Visual summary of the relative errors made by WOC, the three variants of the method in (A) and the center of the clusters obtained at low social weight at four levels of resolution in (B). Data taken from Lorenz et al. [18].

mean of the cluster of people with low social weight is 422, 481, 512 and 491 km for γ_{w_s} = 2, 3, 4 and 6, respectively (Fig 2.5B). In particular, it is not necessary that the value γ_{w_s} chosen for the clustering corresponds with a distribution showing peaks. For example, the distribution with $\gamma_{w_s} = 6$ does not show peaks and it is clustered into approximately the same two clusters than the distribution with $\gamma_{w_s} = 3$ that shows two clear peaks. The values obtained are -42%, -34%, -30% and -33% off the true value of 734 km. The cluster at high

social weight correspond to individuals with larger errors (-69%, -67%, -71% and -67% for γ_{w_s}=2, 3, 4 and 6, respectively). The WOC value is typically a value between the ones at low and at high social weights, here 302, -59% off the true value.

2.2.7 Test of the methods with other questions

So far we have seen that using the individuals with lowest social weight we can estimate 'What is the Swiss/Italian border length?' better than using WOC. The results were robust under changes in the method to extract the individuals with low social weights, with a total of 7 variants of the methods used improving over WOC (Fig 2.5C). We then applied the same methods to the remaining 5 questions from the experiments in Lorenz et al. [18]. We found a subpopulation with a significant resistance to social influence in 3 of the remaining questions (Fig 2.6 and Table 2.2 for a summary.

For the question of 'Number of rapes in 2006 in Switzerland' the geometric mean of individuals of low social weight as measured by Eq 2.9 and its two variants gives the same value as there is a single significative group at a value of 624, much larger than the WOC result of 257 (Fig 2.6A, 'resist 1,2,3'). This corresponds to a much smaller error (-2.3%) than the WOC (-60%) respect to the truth at 639. The distribution of estimations does not show a structure of two peaks separated at low and high social weight (Fig 2.6B, $\gamma_{w_s} = 6,4,3$) and at high resolution there are too many peaks with very few individuals each (Fig 2.6B, $\gamma_{w_s} = 2$) so a method based on peaks is not appropriate for this question.

For the 'Number of assaults in 2006 in Switzerland', the geometric mean in 2.9 and the two variants considered have a large deviation from the WOC value of 3685 to 6654, 6313 and 7557, respectively (Fig 2.6C, 'resist 1','resist 2','resist 3'). They correspond to errors of -28%, -32% and -18%, respectively, much lower than the -60% error of WOC. The clustering method obtains the same value of 7699 for γ_{w_s}=3, 4 and 6 (Fig 2.6D, $\gamma_{w_s} = 6,4,3$) and for $\gamma_{w_s} = 2$ the resolution is too high and reveals at least four peaks with very few individuals per peak (Fig 2.6D, $\gamma_{w_s} = 2$). For γ_{w_s}=3,4, and 6 the error is -17% of the true value 9272 compared to the -60% error of the WOC of 3685.

For the question about the 'Population density of Switzerland' the geometric mean in 2.9 does not find a subpopulation resisting social influence with estimations significantly different to WOC (Fig 2.6E). The clustering method finds for γ_{w_s}=2,3,4 and 6 the values 174, 177, 177 and 171, respectively (Fig 2.6F, $\gamma_{w_s} = 6,4,3,2$). Compared to the true value of 184,

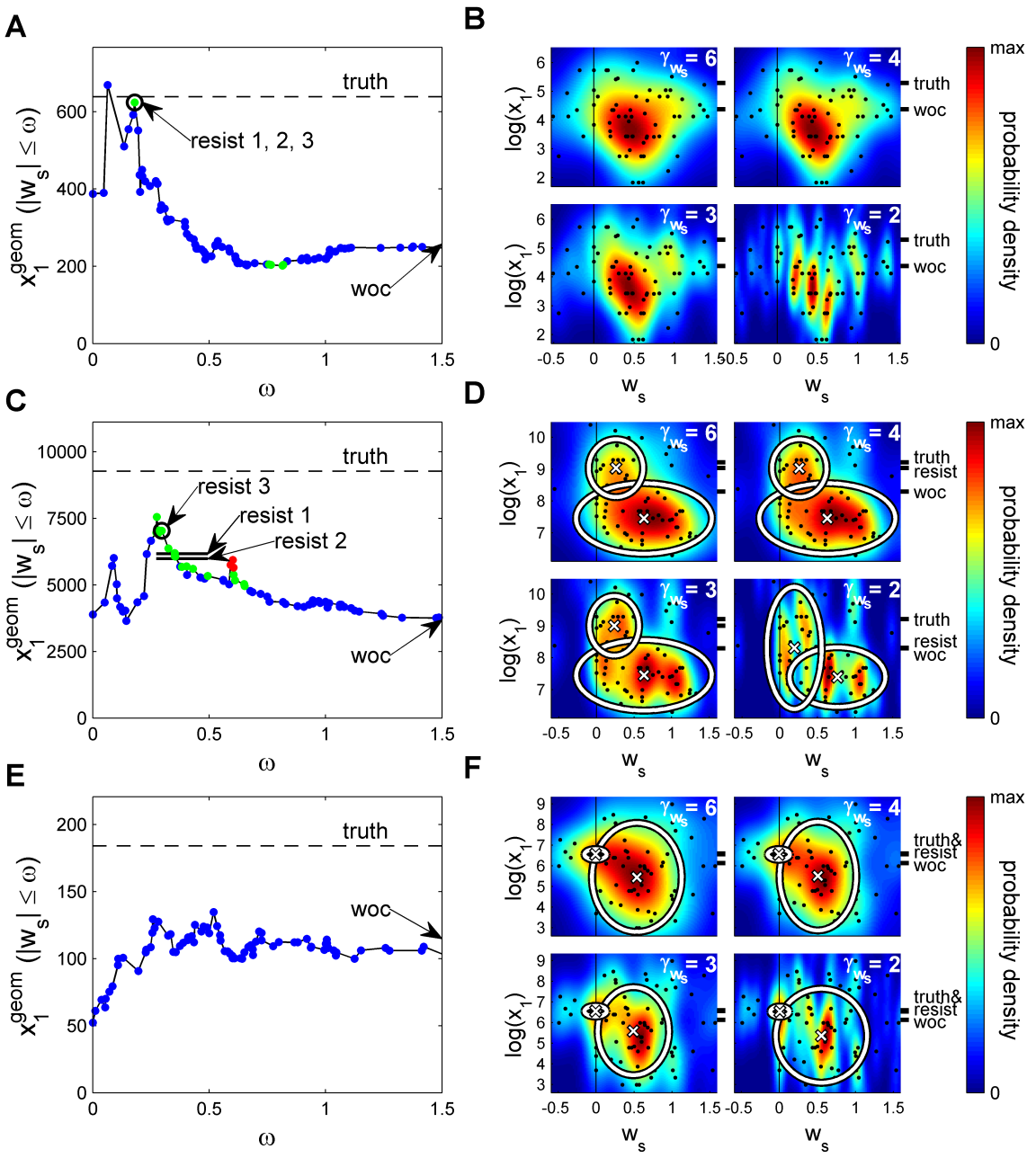

Fig. 2.6 **Wisdom of those resisting social influence for three questions.** Analysis as in Fig 2.2A,B but for the questions **(A, B)** "How many rapes were officially registered in Switzerland in 2006?", **(C, D)** "How many assaults were officially registered in Switzerland in 2006?", and **(E, F)** "What is the population density of Switzerland in inhabitants per square kilometer?" Data taken from Lorenz et al. [18].

Question	truth	WOC	resist 1	resist 2	resist 3	$\gamma_{w_s}=6$	$\gamma_{w_s}=4$	$\gamma_{w_s}=3$	$\gamma_{w_s}=2$
Border	734	302 (-59%)	714 (-2.7%)	689 (-6.2%)	780 (+6.3%)	491 (-33%)	512 (-30%)	481 (-34%)	422 (-42%)
Rapes	639	257 (-60%)	624 (-2.3%)	624 (-2.3%)	624 (-2.3%)	-	-	-	-
Assaults	9272	3685 (-60%)	6170 (-33%)	5984 (-35%)	7037 (-24%)	7699 (-17%)	7699 (-17%)	7699 (-17%)	3881 (-58%)
Population	184	115 (-38%)	-	-	-	171 (-7.3%)	177 (-4.0%)	177 (-4.0%)	174 (-5.7%)

Table 2.2 **Comparison of true value, 'wisdom of the crowds' (WOC) and the prediction from the subgroup of individuals resisting social information. resist 1** computed from individuals with low social weights and contributing more the values of ω with higher significance (Eq 2.9). **resist 2** computed as 'resist 1' but not weighting the different ω differently depending on significance levels. **resist 3** corresponds to the value of ω with highest significance. γ_{w_s}**= 6, 4, 3, 2** give the central values of the peaks at low social weights obtained from a Gaussian mixture at a resolution in the direction of social weight w_s obtained introducing the values of γ_{w_s} in Eq 2.8. **Border**, 'What is length of the Swiss/Italian border?'; **Rapes**, 'How many rapes were officially registered in Switzerland in 2006?'; **Assaults**, 'How many assaults were officially registered in Switzerland in 2006?'; **Population**, 'What is the population density of Switzerland in inhabitants per square kilometer?'

these values are -5.7%, -4.0%, -4.0% and -7.2% off the true value of 184 while the WOC value of 115 is -38% off.

2.2.8 The Wisdom of the Confident

Our analysis shows that estimation is improved when there is a subpopulation significantly resisting social influence, and the estimation of those individuals may be viewed as *wisdom of the confident*. The seven variants of the methods improve upon WOC and in many cases the improvement is very large (Table 2.2). The success of the method rests in the correlation between resistance to social influence and closeness to the true value seen in the data. It is also interesting to consider some properties of the resisting individuals. The proportion of these individuals is $25 \pm 13\%$ using the methods based on Eq 2.9 and and $10 \pm 3\%$ for the methods based on the peaks of the distribution. The individuals that resist social influence are not the same in all questions. We only find a significant overlap between the first and second questions (Fig 2.7A, $p < 0.05$).

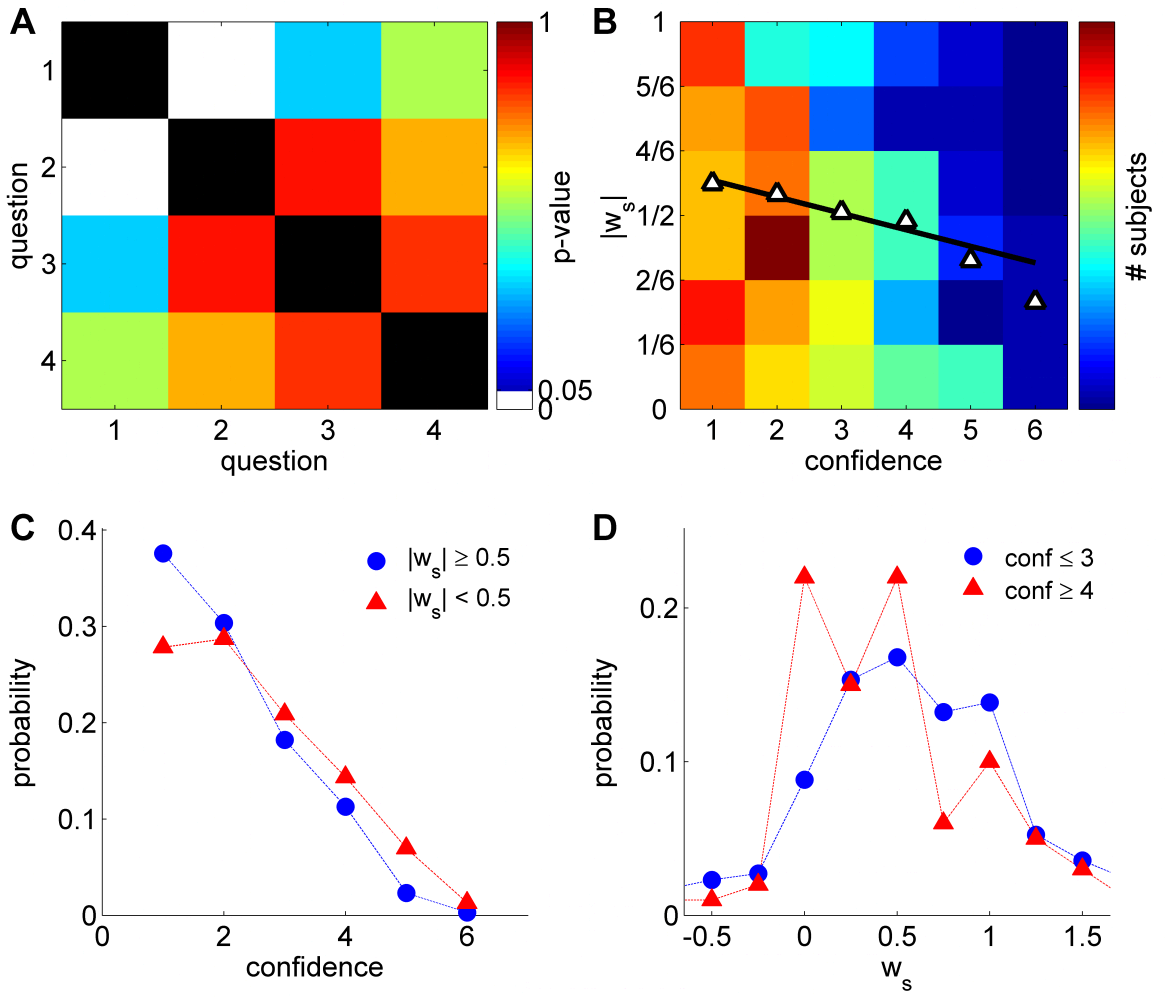

Fig. 2.7 **Characterization of individuals resisting social information.** (**A**) Significance of the coincidence of resisting individuals ($w_s < 0.5$) for every pair of the 4 questions analyzed in main text. There is only a significant overlap of individuals resisting influence for questions 1 and 2 ("What is the length of the border between Switzerland and Italy in kilometers?" and "How many rapes were officially registered in Switzerland in 2006?"). (**B**) Correlation of social weight (only for) and declared confidence is significant ($p < 0.0003$) but weak ($R^2 = 0.03$) respect to linear regression (straight line). Triangles at mean social weight for each confidence value.In colors the joint distribution of social weights and confidence values, showing large dispersion from regression line. (**C**) Probability of the declaration of confidence for individuals resisting (red triangles) and not resisting (blue circles) social influence. (**D**) Probability that an individual has a social weight when they declare a low (blue circles) and high confidence (red triangles). Data taken from Lorenz et al. [18].

2.2.9 Declared confidence

Resistance to social information may be viewed as a behavioral measure of confidence. Its success is not a trivial result as other measures of confidence like declared confidence in a scale from 1 to 6 does not necessarily improve accuracy (Sniezek and Henry [29], Bahrami et al. [2], Koriat [11], Mahmoodi et al. [19]). We thus decided to compare why the two measures give different results. We found a significant but very low correlation between resistance to social information and declared confidence (Fig 2.7B, $p < 0.001$, $R^2 = 0.03$). While there are approximately equal numbers of resisting and non-resisting individuals (Fig 2.3B), most of the population declares low values of confidence, even the majority of those resisting social influence (Fig 2.7C, triangles). Individuals declaring high values of confidence (Fig 2.7D, triangles), in general resist social influence more than those with low values, but a relevant proportion does not resist social influence. The two measures are correlated but are very different and it is then unsurprising than a method like the one proposed here for social resistance does not work for declared confidence (Fig 2.8).

2.3 Discussion

We have here proposed to extract information from the collective using those individuals resisting social influence. The methods proposed extract the information a collective considers of high private quality. We obtained better collective estimations than the 'wisdom of crowds' (Galton [7], Surowiecki [30], Page [23], Lee and Shi [15], Wagner and Vinaimont [32], Easley and Kleinberg [6], Krause et al. [12], King et al. [9], Lorenz et al. [18]) using the data from Lorenz et al. [18], especially for cases in which the crowd shows a very large bias. The methods work because resistance to social influence correlates with closeness to the true value. The correlation does not need to be very strong, that is, we do not need experts (Wolfers and Zitzewitz [34], Whitehill et al. [33], Lee et al. [16]). Instead, we use the geometric mean of those individuals that get influenced less by social information and this group can still show a large standard deviation.

We used two types of methods. One based on 2.9, taking all individuals below a value of social weight that give a result different from WOC. This method gave predictions very close to true values for those cases in which the joint distribution of estimations and social weight does not show a complex structure at low social weights. When this method does not give significant results, one can resort to a method based on clustering in the space

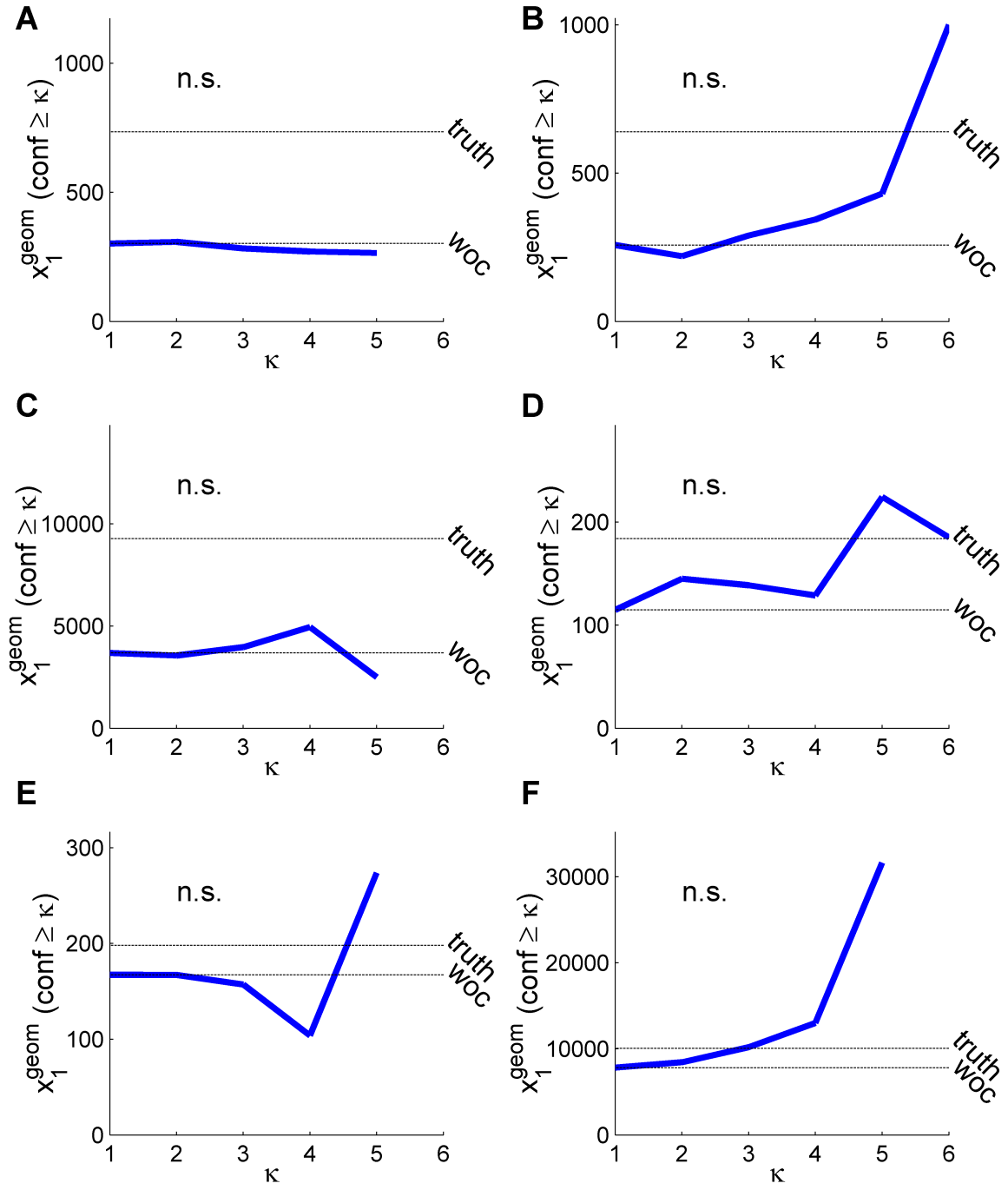

Fig. 2.8 **Collective estimations for individuals declaring confidence.** We used a method analogous to that of Figs 2.5A and 2.6A,C,E in main text but for declared confidence instead of social weight. Geometric mean of individuals declaring a value of confidence ($conf$) in their estimation higher or equal than an integer κ. No value is found to be significant ($p_{min} > 0.08, \bar{p} = 0.48$). Question wordings and labels as in Section 2.4.1. Data taken from Lorenz et al. [18].

defined by estimations and social weights. This second type of methods take into account less individuals, but we found they improve upon WOC. The two methods together can be used to understand the relevant subjects in the estimation. For example, Eq 2.9 does not give significant results for the question on the '*Population density of Switzerland*' (Fig 2.6E). Inspection of the density shows that while there is a strong peak at low social weight with an estimation very different from WOC (Fig 2.6F), there are individuals giving much lower estimations and thus making the geometric mean of individuals with low social weight not different from WOC.

Our proposal makes use of individuality to improve upon WOC. It is interesting to speculate what type of individuality is most compatible with our results. One type of individuality would simply be that all individuals use a similar procedure to answer a question but their levels of noise are different. One way to model this would be to extend our models to incorporate that individuals are most likely to give the correct answer but they have different levels of noise (Section 2.4.10). This model gives very poor predictions (Table 2.3). The reason is that the data seems more compatible with different subgroups of people with different biases from the truth, for example the low and high peaks in the joint density in Fig 2.4A. This can be modeled in that the most probable estimation is shifted away for the true value with different biases in different individuals. As biases are defined respect to truth, this extension would not be predictive. Instead, we propose the methods in the main text, by which we extract the subgroup of individuals of low social weight as the more accurate ones on average.

The idea that different individuals or subgroups of individuals have different biases is compatible with the existence in the population of different procedures to solve a problem, each of them with a different bias. According to this view, a possible origin of the data for the question about the Swiss/Italian border as an example could be the following. This question might be answered estimating the approximate length of a straight line separating the two countries, which is 288 km as measured from a map in http://www.freemaptools.com/measure-distance.htm. Interestingly, the cluster of individuals with highest social weight is characterized by an estimation of 216 ± 157 km compatible with these very low values. A procedure more sophisticated than simply the length of a straight line consists in using the shape of the border. Another procedure is to use memorized data to retrieve its value. The cluster at low social weight is characterized by an estimation of 512 ± 269 km and the geometric mean at low social weights by values in the interval 650-800 km, compatible with these more

sophisticated procedures. This idea of different procedures might also explain the different susceptibilities to social information. Those individuals using the shape of the Swiss/Italian border would in general not consider as very important social information with values so much lower than their estimations. This is because these values would be incompatible with the shape, for example values closer to a straight line. In contrast, individuals using a straight line approach might be willing to consider higher values, as they might have only taken this approach as a very rough approximation they could make because they had difficulties finding how to estimate the full shape. All individuals might declare low confidence levels as they can be very noisy within their approach, but they might still consider differently values more compatible with other approaches.

A second and complementary explanation of individuality is that individuals have different levels of expertise on the subject or even in general exercises of estimation. This level of expertise is probably not high enough for the individuals to declare it, but it would be enough to act upon it when confronted with social influence.

The methods proposed to improve upon WOC do not correspond to a common situation in which humans interact naturally. Instead, it is a protocol that can be used to extract high quality information in human collectives even if it is present only in a minority of the group. Its value relies on improving upon WOC by eliminating the people that are not confident in their private estimations. And using how much each individual is influenced by others as a measure of confidence seems to extract the correct individuals, unlike methods based on declared confidence (Sniezek and Henry [29], Bahrami et al. [2], Koriat [11], Mahmoodi et al. [19]). Our results point to measures of confidence not based on declaration as a means to gather high quality private information in a group. Response time, perseverance or pay-offs in decision systems might be implementations to test experimentally. An open problem is in which circumstances social influence or these other measures of confidence can be used by humans to improve individual and collective decisions in naturalistic settings.

2.4 Materials and Methods

2.4.1 Data

We tested the model by reanalyzing a dataset in which subjects made estimations before and after social influence (Lorenz et al. [18], data can be downloaded from http://www.pnas.org/content/108/22/9020?tab=ds). This is a rich dataset that can be used as a reference to

test models of social influence (Mavrodiev et al. [21]). In these experiments subjects were asked to privately estimate the answer to six questions (Lorenz et al. [18]): (A) 'What is the length of the border between Switzerland and Italy in kilometers?', (B) 'How many rapes were officially registered in Switzerland in 2006?', (C) 'How many assaults were officially registered in Switzerland in 2006?', (D) 'What is the population density of Switzerland in inhabitants per square kilometer?', (E) 'How many murders were officially registered in Switzerland in 2006?' and (F) 'How many more inhabitants did Zurich gain in 2006?' After their private estimation for each question, each subject could receive social information consisting in either receiving on a computer screen a diagram depicting the private estimates of each member of the group ('full information' condition) or more simply their arithmetic mean ('aggregated information' condition). To test that the observed effects were due to social interactions, they also used control groups that also estimated twice but without social influence in between ('no information' condition). The experimental data was obtained using 144 people organized in 12 groups of 12 people. Each group was asked 6 questions, 2 in each of the three conditions. For each question, the process of receiving social information and revising the estimation was repeated four times, so subjects provided five estimates in a row. However, we only analyzed the change in opinion from the first reply to the second. An iterative model, where the social information received in a previous trial would be included in the private information in the current, would be suitable for analyzing subsequent trials.

2.4.2 Derivation of Eq 2.1

We have shown elsewhere that choices in animal collectives are well described using estimation theory (Pérez-Escudero and de Polavieja [26], Arganda et al. [1]). These models are based on subjects using the probability that Y is the best option, which using Bayes theorem might be written as (Pérez-Escudero and de Polavieja [26])

$$P(Y \text{ is best option}|B,C) \propto P(B|Y \text{ is best option},C)\,P(Y \text{ is best option}|C)\,, \qquad (2.10)$$

where C is the private information the individual has, and B the observed behaviors of the other subjects, specifically how many of them had chosen each of the two options X and Y. The main idea of these models is that animals estimate using both private and social information, and these two sources enter in the estimation as a multiplication. We now obtain a form of this rule when estimating the value of a continuous variable x. The distribution of

estimations made by humans is a log-normal (Lorenz et al. [18]). For this reason, we here use the variable $y \equiv \log(x)$, so a log-normal distribution in x is a normal distribution in y. An estimation that y is the correct value based only on private information (p) would then be modeled as

$$f_Y\left(y \text{ is correct}|p\right) = \frac{1}{\sigma_p\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu_p}{\sigma_p}\right)^2} = N\left(\mu_p, \sigma_p\right). \tag{2.11}$$

This expression is simply saying that based on private information the individual estimates that the correct value of y has a probability centered at μ_p with standard deviation σ_p.

More generally, individuals make estimations using private and social information. We are here interested in the case in which the social information is made of the estimations made by other individuals, $\vec{y}$. The estimating individual would then compute the probability that y is the correct value given the private information (p) and the estimations by others $\vec{y}$, which by Bayes theorem can be expressed as (Papoulis and Pillai [24])

$$f_Y\left(y \text{ is correct}|p,\vec{y}\right) = \frac{f_{\vec{Y}}\left(\vec{y}|y \text{ is correct}, p\right) f_Y\left(y \text{ is correct}|p\right)}{\int_{-\infty}^{\infty} f_{\vec{Y}}\left(\vec{y}|y \text{ is correct}, p\right) f_Y\left(y \text{ is correct}|p\right) dy}, \tag{2.12}$$

where $f_Y\left(y \text{ is correct}|p\right)$ is in Eq 2.11. The term $f_{\vec{Y}}\left(\vec{y}|y \text{ is correct}, p\right)$ is the probability that the other individuals give the estimations $\vec{y}$ when y is the correct value. It is thus a measure of how reliable the other individuals are. We here consider the cases in which the estimations of the others were given independently of each other as

$$f_{\vec{Y}}\left(\vec{y}|y \text{ is correct}, p\right) = n! \prod_{i=1}^{n} f_{Y_i}\left(y_i|y \text{ is correct}, p\right). \tag{2.13}$$

The term $n!$ counts all the possible sequences of decisions that lead to the set of values $\vec{y}$, as we are not interested in which particular subject emitted the particular estimate y_i. We model the terms $f_{Y_i}\left(y_i|y \text{ is correct}, p\right)$ also as Gaussians of the form

$$f_{Y_i}\left(y_i|y \text{ is correct}, p\right) = \frac{1}{\sigma_s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-y}{\sigma_s}\right)^2} = N\left(y, \sigma_s\right). \tag{2.14}$$

This expression means that, when y is the correct value, individual i is modeled as being able to give this value with the highest probability but with standard deviation σ_s. This term thus measures how reliable each individual i is, and assumes the same reliability for all individuals and no bias. To add that each of the other individuals has a different reliability we would

have a different standard deviation for each individual, $\sigma_{s,i}$. To add a global bias we would have $(y_i - y - a)^2$ instead of $(y_i - y)^2$ in Eq 2.14 or individual bias as $(y_i - y - a_i)^2$.

Using Eq 2.14, Eq 2.13 can be written as

$$f_{\vec{Y}}\left(\vec{y}|y \text{ is correct}, p\right) = n! \prod_{i=1}^{n} \frac{1}{\sigma_s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - y}{\sigma_s}\right)^2} = N\left(\mu_s, \sigma_s\right), \tag{2.15}$$

where

$$\mu_s \equiv \frac{1}{n}\sum_{i=1}^{n} y_i = \log\left(\left[\prod_{i=1}^{n} x_i\right]^{1/n}\right) \tag{2.16}$$

is the logarithm of the geometric mean of the estimates made by others, and B is a term that does not depend on y and that cancels out in the next step. Substituting Eq 2.15 and Eq 2.11 into Eq 2.12, we obtain

$$f_Y\left(y \text{ is correct}|p, \vec{y}\right) = \frac{N\left(\mu_p, \sigma_p\right) N\left(\mu_s, \sigma_s\right)}{N\left(\mu_p, \sigma_p\right) N\left(\mu_s, \sigma_s\right)} = N\left(\mu_f, \sigma_f\right), \tag{2.17}$$

with

$$\mu_f \equiv \frac{\sigma_s^2}{\sigma_s^2 + n\sigma_p^2}\mu_p + \frac{n\sigma_p^2}{\sigma_s^2 + n\sigma_p^2}\mu_s, \quad \sigma_f \equiv \frac{\sigma_s\sigma_p}{\sqrt{\sigma_s^2 + n\sigma_p^2}}. \tag{2.18}$$

A more compact notation is obtained defining a "private weight" w_p and "social weight" w_s as

$$w_p \equiv \frac{\sigma_s^2}{\sigma_s^2 + n\sigma_p^2}, \quad w_s \equiv \frac{n\sigma_p^2}{\sigma_s^2 + n\sigma_p^2}, \tag{2.19}$$

so the parameters in Eq 2.18 can be expressed as

$$\mu_f = w_p\mu_p + w_s\mu_s, \quad \sigma_f = \sqrt{1 - w_s}\sigma_p. \tag{2.20}$$

Introducing Eq 2.20 into Eq 2.17, the probability distribution of the logarithm of estimations when subjects combine their private information p and the estimates by the others $\vec{y}$ is written in the compact form

$$f_Y\left(y \text{ is correct}|p, \vec{y}\right) = \frac{1}{\sqrt{1 - w_s}\sigma_p\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y - \left(w_p\mu_p + w_s\mu_s\right)}{\sqrt{1 - w_s}\sigma_p}\right)^2}. \tag{2.21}$$

This is an expression that would model the estimation step for each individual, and from this each individual would produce a concrete value. There are several decision rules individuals could be applying. A simple deterministic rule would simply consist in choosing the value with highest probability. This is however an unlikely rule given the many possible sources of noise, for example memory noise (Vul and Pashler [31]). We will adopt here probabilistic matching, as we have done previously for data in other species (Pérez-Escudero and de Polavieja [26], Arganda et al. [1]). This is a probabilistic rule that does not add additional parameters in the model, according to which the probability of choosing a value y is simply the probability that y value is the correct one, that is, $f_Y\left(y \text{ is correct}|p,\vec{y}\right)$ in Eq 2.21.

2.4.3 Derivation that $\mu_s = \mu_p$ and $\mu_s = \mu_p + \sigma_p^2/2$

We have shown in Eq 2.16 that the social term μ_s reduces to the logarithm of the geometric mean of the estimations made by others, $\mu_s = \log(x_s)$, with

$$x_s = \left(\prod_{i=1}^{n} x_i\right)^{1/n}. \tag{2.22}$$

But the geometric mean x_s is an estimator of the median $\exp(\mu_p)$ of the population (Parkin and Robinson [25], Limpert et al. [17]), and consequently

$$E[\mu_s] = \mu_p \rightarrow E\left[\mu_f\right] = w_p\mu_p + w_s E[\mu_s] = w_p\mu_p + w_s\mu_p = \mu_p, \tag{2.23}$$

making the final distribution in Eq 2.21 to be centered at μ_p. In a second type of experiments that we consider (Lorenz et al. [18]), the social information is not the set of estimations made by all other subjects but simply the mean of all of them. When subjects treat this social information in the same way they treat a set of estimations made by other subjects, then $\exp(\mu_p)$ is the arithmetic mean,

$$x_s = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.24}$$

For a log-normal distribution, the expected value of the mean is of the form (Johnson et al. [8])

$$E[x_s] = \exp\left(\mu_p + \frac{\sigma_p^2}{2}\right). \tag{2.25}$$

In this case the mean of the final distribution in Eq 2.21 is

$$E\left[\mu_f\right] = w_p\mu_p + w_s E\left[\mu_s\right] = w_p\mu_p + w_s\left(\mu_p + \frac{\sigma_p^2}{2}\right) = \mu_p + w_s\frac{\sigma_p^2}{2}. \tag{2.26}$$

We will then use Eq 2.21 as the distribution of estimations after social interactions with Eq 2.23 when the social interactions are all the estimations of the other subjects ("full information" condition in the main text) and Eq 2.26 when they receive the mean value of the other subjects ("aggregated information" condition).

2.4.4 Version of Eq 2.1 used for z-score

We also used in the main text a z-score instead of the variable $y = \log(x)$ for Fig 2.1. When the distribution before social interactions is a log-normal with parameters

$$\mu_o = \mu_p, \quad \sigma_o = \sigma_p, \tag{2.27}$$

for the z-score

$$z \equiv \frac{\log\left(x - \mu_p\right)}{\sigma_p} \tag{2.28}$$

the distribution has parameters

$$\mu_{zo} = 0, \quad \sigma_{zo} = 1. \tag{2.29}$$

After social interactions in the "full information" condition the final distribution Eq 2.21 has the same mean (Eq 2.23) and a reduced standard deviation

$$\mu_f = \mu_p, \quad \sigma_f = \sqrt{1 - w_s}\sigma_p, \tag{2.30}$$

that in the z-score gives a distribution with parameters

$$\mu_{zf} = 0, \quad \sigma_{zf} = \sqrt{1 - w_s}. \tag{2.31}$$

When the social interaction consists in giving the arithmetic mean ("aggregated information" condition) the final distribution Eq 2.21 has a different mean (Eq 2.26) and a reduced standard deviation

$$\mu_f = \mu_p + w_s\frac{\sigma_p^2}{2}, \quad \sigma_f = \sqrt{1 - w_s}\sigma_p, \tag{2.32}$$

that in the z-score corresponds to a Gaussian with parameters

$$\mu_{zf} = \frac{\mu_p + w_s \frac{\sigma_p^2}{2} - \mu_p}{\sigma_p} = w_s \frac{\sigma_p}{2}, \quad \sigma_{zf} = \sqrt{1 - w_s}. \tag{2.33}$$

2.4.5 Significance tests used for the difference of means or variances

A complete list of significance tests can be found in Table 2.1. In the main text, unless otherwise stated, we computed p-values explicitly without assumptions about the data as the probability that the experimental result is obtained at random. For example, to find whether two distributions have a significantly different value of some parameter θ (in our case, the mean or the variance), we performed a permutations method. We mixed the two samples and randomly divided the resulting set into two subsets. Then, we computed the sample value of the parameter in each of the subsets and extracted the difference $d \equiv |\theta_1 - \theta_2|$. We repeated this process 10^6 times, obtaining a distribution of differences d. The significance p is the proportion of d values bigger than the difference of the parameters between the two original samples.

2.4.6 Smoothing of distributions

The distributions were calculated using Gaussian kernel smoothing (Silverman [28]). The 1D version of Gaussian kernel smoothing was applied for social weights w_s in Fig 2.3B as (Silverman [28])

$$f(w_s) = \frac{1}{\sqrt{2\pi}\sigma n} \sum_{i=1}^{n} \exp\left(-\frac{(w_s - w_{s,i})^2}{2\sigma^2} \right), \tag{2.34}$$

with $\{w_{s,i}\}$ the values of the social weights obtained from experiments using Eq 2.7, n the length of the sample and $\sigma \equiv \hat{\sigma} n^{-1/\gamma}$ the bandwidth with $\hat{\sigma}$ the standard deviation of the sample and γ the resolution coefficient. We set the resolution coefficient to $\gamma = 5/2$ half its optimal value (Silverman [28]), a value that allows the visualization of the main structure of the distribution. We were interested in the interval [0,1] and did not then consider points outside (-1,2) in our calculations of the bandwidth, avoiding tail effects. The 2D case of Gaussian kernel smoothing is described in the main text, Eq 2.8.

2.4.7 Significance for the geometric mean method

To find whether the group of individuals with $w_s < \omega$ in Figs 2.5A and 2.6A,C,E has geometric mean significantly different from WOC, we used the following procedure. Each ω corresponds to a subgroup of n_ω individuals. We obtained 10^5 random sets of n_ω estimations from the whole crowd and computed the geometric mean of each set, g. The significance of $x_1^{geom}(w_s \leq \omega)$ is the proportion of values of g at least as far to the wisdom of the crowd (geometric mean) as $x_1^{geom}(w_s \leq \omega)$.

2.4.8 Significance test used for the method using the distributions

To divide the region of maximum density into two clusters, we performed an Expectation Maximization (EM) algorithm to obtain a mixture of two Gaussians (McLachlan and Peel [22], Bilmes and others [3]). More specifically, for each value of γ_{w_s} we selected those individuals whose social weight and estimate $(w_{s,i}, \log x_i)$ lied in the zone of maximum probability, defined as that where the probability in Eq 2.8 is at least equal than half of the maximum. Then an EM algorithm was applied to the selected data points to find the maximum likelihood estimates of the parameters of a Gaussian mixture with two components.

2.4.9 Significance test of whether two questions share the same resisting individuals

To find whether two questions shared a significant number of individuals with low $|w_s|$, we used the exact expression for the probability that two samples from a finite population have a certain number of elements in common. Specifically, we want to compare two selections at random of N and M subjects from a group of n subjects. Both selections are made from the entire original group, so they may have common elements. We are interested in the probability that the two selections have Z or more subjects in common. The probability that in the group of subjects (the "M-group") you have exactly Z of those in the group of N subjects (the "N-group") is the ratio of the number of favorable cases and all the possible results. The number of favorable cases is given by the product of

$\binom{N}{Z}$: number of combinations of N elements taken in groups of Z. Once the N-group is fixed, the above number counts all the possible groups of Z elements that can be extracted from it.

$\binom{n-N}{M-Z}$: number of combinations of the other n_N elements, taken in groups of $M-Z$. This number counts all the possible ways to complete the M-group once the Z common elements are fixed, but without selecting any more elements from the N-group.

The total number of cases is

$\binom{n}{M}$: combinations of n elements taken in groups of M. This is the number of all possible M-groups that can be formed from the entire original group.

Then the probability that the two selections have Z subjects in common is

$$P(Z) = \frac{\binom{N}{Z}\binom{n-N}{M-Z}}{\binom{n}{M}}. \tag{2.35}$$

Note that the probability is symmetric under the interchange of N and M:

$$P(Z) = \frac{\binom{M}{Z}\binom{n-M}{N-Z}}{\binom{n}{N}}. \tag{2.36}$$

The p-value, defined as the probability of having Z or more common elements is

$$p \equiv P(\xi \geq Z) \sum_{\xi=Z}^{min(M,N)} = \frac{\binom{M}{\xi}\binom{n-M}{N-\xi}}{\binom{n}{N}}. \tag{2.37}$$

2.4.10 Bayesian weights as a simple model of individuality

In this section we derive a model like that of Section 2.4.2 but introducing individuality. In section 2.4.2, we assumed that the estimating individual models all the other individuals as giving the correct value with the highest probability and with the same standard deviation. In this section we will assume that the standard deviation is different for each individual. The interest of this approach is that we could in principle use it not only as a model of how human subjects react to social influence, but also as a procedure to get a good prediction for each question in the main text using the individuality of the estimations. The model results in a prediction given by the weighted average of the estimations of the population, with each individual weighting more the lower the social weight (Eq 2.41 below). This result is intuitive with individuals having more importance in the prediction the less they are influenced by social information. However, we also show in this section that this approach does not give good predictions. The reason for this failure was expected as we are only modelling individuality in the standard deviation. We are still assuming in this model that all

individuals can give the correct value with the highest probability. The experimental data indicates this is not the case as different individuals or subpopulations can have different biases from the truth (for example, see Fig 2.5B). Modeling individuality in the bias means that each individual can have the highest probability of giving an estimation shifted from the correct value. This can be modeled formally, but as bias is defined respect to truth, the model would have no predictive value. Instead, our proposal is to follow the procedure of the main text, which uses only individuals with low social weight as they are on average closer to correct values. Alternatively, one may consider the methods in the main text as a strong prior and include the weights we obtain here of the individuals extracted.

The derivation of this model is as follows. We introduce individuality by considering that instead of all individuals having the same reliability as in Eq 2.15, each individual has a different value of reliability as

$$n!\prod_{i=1}^{n} f_{Y_i}(y_i|y,p) = n!\prod_{i=1}^{n}\frac{1}{\sigma_{s,i}\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y_i-y}{\sigma_{s,i}}\right)^2} = BN(\mu_s,\sigma_s) \tag{2.38}$$

with individuals with smaller standard deviation $\sigma_{s,i}$ more reliably giving an estimation closer to the correct value, and we can use the properties of the product of Gaussians (Bromiley [4]) to write

$$\mu_s = \frac{\sum_{i=1}^{n}\left(y_i/\sigma_{s,i}^2\right)}{\sum_{i=1}^{n}\left(1/\sigma_{s,i}^2\right)}, \quad 1/\sigma_s^2 = \sum_{i=1}^{n}\left(1/\sigma_{s,i}^2\right). \tag{2.39}$$

We can estimate the individual standard deviations $\sigma_{s,i}$ in the following way. The $\sigma_{s,i}$ measures the width of the probability distribution for each subject, a role played by σ_p in Eq 2.11. Similarly to that case, we can use how each subject reacts to social influence, Eq 2.20, to obtain how standard deviation relates to the social weight as

$$\sigma_{s,i} = \frac{\sigma_f}{\sqrt{1-w_{s,i}}}, \tag{2.40}$$

with $w_{s,i}$ the experimental social weight. We assume that σ_f is the same for all subjects, as an approximation we need to take with the data at hand. The specific value of σ_f is not

Question	truth	WOC	ignore $\|w_s\| > 1$	ignore $w_s > 1$, and (collapse $w_s < 0$ to 0)	ignore $\|w_s\| > 1$ and $\|w_s\| < 0$
Border	734	302	397	389	333
		(-59%)	(-46%)	(-47%)	(-55%)
Rapes	639	257	241	256	244
		(-60%)	(-62%)	(-60%)	(-62%)
Assaults	9272	3685	4834	4823	4412
		(-60%)	(-48%)	(-48%)	(-52%)
Population	184	115	110	116	100
		(-38%)	(-40%)	(-37%)	(-46%)
Murders	198	167	149	153	139
		(-16%)	(-25%)	(-23%)	(-30%)
Immigrants	10067	7819	8414	9224	7666
		(-22%)	(-16%)	(-8.4%)	(-24%)
	Average	(-42%)	(-40%)	(-37%)	(-45%)

Table 2.3 **Prediction of the Bayesian Weights Model. Border**, "What is length of the Swiss/Italian border?"; **Rapes**, "How many rapes were officially registered in Switzerland in 2006?"; **Assaults**, "How many assaults were officially registered in Switzerland in 2006?"; **Population**, "What is the population density of Switzerland in inhabitants per square kilometer?"; **Murders**, "How many murders were officially registered in Switzerland in 2006?"; **Immigrants**, "How many more inhabitants did Zurich gain in 2006?"

important as it cancels out in the next steps. Using Eq 2.40, we can express Eq 2.39 as

$$\mu_s = \frac{\sum_{i=1}^{n} \left(\frac{1-w_{s,i}}{\sigma_f^2} y_i \right)}{\sum_{i=1}^{n} \left(\frac{1-w_{s,i}}{\sigma_f^2} \right)} = \frac{\sum_{i=1}^{n} \left[(1-w_{s,i})\, y_i \right]}{\sum_{i=1}^{n} (1-w_{s,i})}. \tag{2.41}$$

Expression 2.41 gives the prediction of this model as a weighted average of the logarithm of estimations, with weights given by $1 - w_{s,i}$, and $w_{s,i}$ the social weights obtained from data. Those individuals that are influenced more by social information weight less in the prediction, as expected.

We compared the prediction of Eq 2.41 with the correct values for the 6 experimental questions in the data (Lorenz et al. [18]). The experimental values of $w_{s,i}$ can be below 0 and above 1 unlike those of the theory, so we make the comparison eliminating these cases in different ways, as in the table below. Irrespective of the method, the comparison of prediction and correct values is poor.

These poor predictions were expected as the model only takes into account individuality in the standard deviation of estimations for each individual and not in the biases respect to truth that individuals can have.

Chapter 3

Aggregation Rules in Consensus Decision-Making

3.1 Introduction

In the previous chapter we showed how Bayesian estimation can be used to prove that the geometric mean is the optimal aggregation strategy for certain estimation tasks. The model presented is in principle applicable to situations in which subjects are allowed to interact. In such a situation the social weight w_s becomes more difficult to compute and even to conceptualize. Nevertheless, there can be experimental situations in which subjects give an estimate of the unknown quantity before discussing with other individuals. At least what could be checked is whether the integration of opinions of various individuals is made in an optimal way. Optimal is defined in the sense that, if the estimates before discussion are expected to follow a normal (log-normal) distribution, the arithmetic (geometric) mean is the optimal average under a Bayesian model of integration of private and social information.

In the previous chapter we did not use weights to average the opinions of the group, but used a classifying measure (social weight w_s to select only the subjects we hypothesized had a private information of higher quality. But once the selection was done, the average of the group was applied, as the model predicted it to be the bayes-optimal way to aggregate information from the collective. However, it might happen that, due to the idiosyncrasy of the group that is aggregating the information, another simple strategies may provide a better approximation to the vakue under estimation. By simple strategies we mean here easy to find

or compute extractions of a consensus decision, like would be taking the median value or the mean of the the two more extreme estimates.

If the geometric (or arithmetic) mean was used by all subjects, or a weighted average with all subjects using the same weight distribution for his opinion and the others, all groups would arrive to a consensus decision made of the geometric (or arithmetic) mean of the individual estimates. However, deviations from the average have been widely reported. There can be many causes for deviations from the average, like differences in personality traits of the group members, statistical noise even within each subject, the groups detecting that the true values was far from the average, or being able to distinguish the experts within the group.

3.2 Results

3.2.1 Weighted average of opinions

In the previous chapter we have shown evidence supporting the idea that subjects integrate social information in a Bayesian optimal way with their private information, to update their initial opinions. The outcome of this processing is a weighted geometric mean of the previous individual estimate and the social information coming from other subjects. So we first hypothesized that when reaching a consensus, groups of subjects may be performing noisy versions of the geometric mean of all the n previous individual estimates:

$$c = \left(\prod_{i=1}^{n} x_i \right)^{1/n}, \tag{3.1}$$

with c the consensus or group discussion estimate. In the previous chapter we have shown than the geometric mean arises naturally from a model that assumes log-normal distributions, which we found to be followed by the experimental set of estimates. In the data set analyzed in this chapter the distributions of estimates are right-skewed but were found to follow log-normal distributions only approximately. Therefore, we also explored noisy versions of the arithmetic mean:

$$c = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{3.2}$$

as it is the most intuitive averaging measure and is the best estimator of centrality in several distributions.

As we are suggesting in the previous paragraphs, deviations from the geometric (or arithmetic) mean can be analyzed as pure statistical noise. However, motivated by the weighted geometric mean model we have presented in the previous chapter, we first tried to model deviations as due to different strengths of the members of the group during the discussion process. This can be modeled via a weighted geometric mean of the pre-discussion estimates, with weights bigger the more predominant the subject within the discussion:

$$c = \prod_{i=1}^{n} x_i^{w_i}, \tag{3.3}$$

with $\sum_{i=1}^{n} w_i = 1$. Also a weighted arithmetic mean can be hypothesized:

$$c = \sum_{i=1}^{n} w_i x_i. \tag{3.4}$$

The problem for sets of three or more individuals is that this provides infinite solutions for any given value of the consensus, even imposing that $0 \leq w_i \leq 1$ for every w_i (Fig 3.1). Each of those solutions may reflect different approaches to the process of discussing, such as a tendency to maximize weights, even when some subjects dominate, or a tendency of the more intermediate estimators to yield to the others.

3.2.2 Even weights for the two closer estimates

To overcome the infinite solutions problem presented in the previous section, we explored different models. We then applied the models to the data to find what are the weights and parameters that provide a best representation of the experimental results. We started with a straightforward simplification of the previous model, in which the two closer estimators (denoted x_1 and x_2) are given the same weight:

$$c = (x_1 x_2)^{(1-w)/2} x_3^w, \tag{3.5}$$

for a weighted geometric mean model, and

$$c = (1-w)\frac{x_1 + x_2}{2} + w x_3 \tag{3.6}$$

for a weighted arithmetic mean model. Intuition may lead to the conclusion that the two closer subjects will unite forces, an thus it will be $w < 1/3$. Testing with the same average

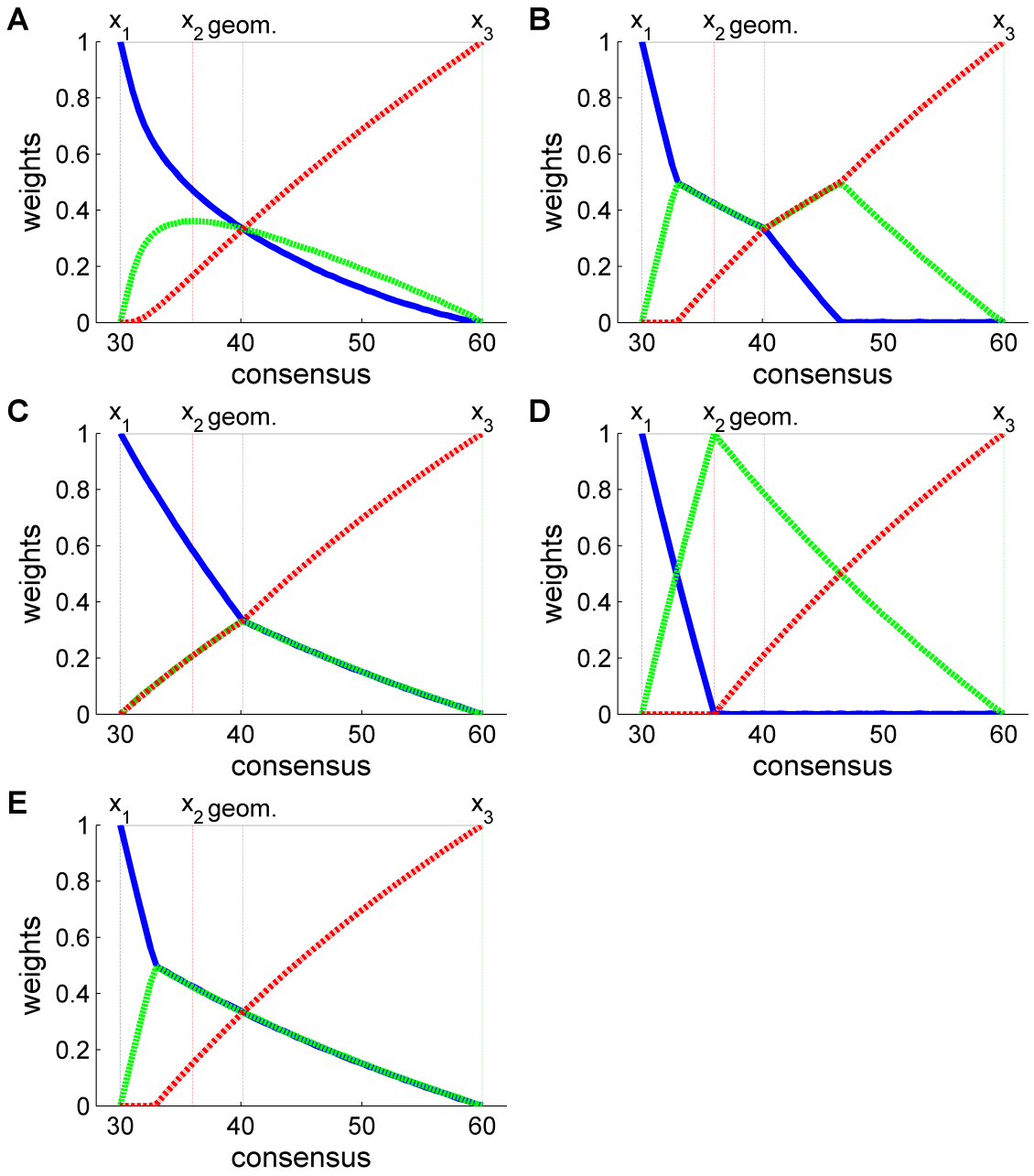

Fig. 3.1 **Different conditions provide different solutions for the weights problem.** For three pre-consensus estimates x_1, x_2, x_3, the weights that provide each consensus value between the lower and higher estimates (blue line for the weight of the subject that gave estimate x_1, green for x_2 and red for x_3). Weights fulfill Eq 3.3, and additional conditions: (**A**) Solution that minimizes the sum of the three weights to the power of $1+\varepsilon$, with $\varepsilon \leqslant 1$. (**B**) Solution that minimizes the maximum of the three weights (equivalent to minimize the sum of weights to a very big power). (**C**) Solution that maximizes the minimum of the three weights. (**D**) Solution that maximizes the weight of the intermediate value. (**E**) Solution that minimizes the sum of the pairwise distance between the weights.

value of w for all 49 groups (Fig 3.2A), the minimum average absolute distance of predicted to real consensus is 4.47, and occurs at $w = 0.31$ for the weighted geometric mean, whereas is 4.80, at $w = 0.21$, for the weighted arithmetic mean model. This shows little improvement from the geometric mean model with even weights for the three subjects. Although the improvement in the case of arithmetic mean is higher, the final prediction is still worse than the predictions from geometric mean. What is more striking is that the weighted geometric mean model is much closer to even weights ($w = 1/3$) than the weighted arithmetic mean model.

The real distribution of weights w_i applied by each group (Fig 3.2B,C), computed for the cases of geometric and arithmetic mean as

$$w_i = \frac{\log(c_i) - \log\left([x_1 x_2]^{1/2}\right)}{\log(x_3) - \log\left([x_1 x_2]^{1/2}\right)} \tag{3.7}$$

and

$$w_i = \frac{c_i - \frac{x_1+x_2}{2}}{x_3 - \frac{x_1+x_2}{2}} \tag{3.8}$$

has mean 0.36 and standard deviation 0.39 for the weighted geometric mean, and mean 0.34 and standard deviation 0.40 for the weighted arithmetic mean.

3.2.3 Spring model

A more complex model is a weighted average of the intermediate estimation and the estimation further to it. The effect of the estimation closer to the intermediate one is increasing the weight of the latter, more the closer they are (Fig 3.3):

$$c = (1 - w)x_2 + wx_3, \tag{3.9}$$

with the weight expressed as a function of two parameters and the distance between the intermediate estimate x_2 and the one closer to it x_1:

$$w = Ae^{-\alpha|x_2 - x_1|}. \tag{3.10}$$

The best average prediction of consensus happens at the values of the parameters $A = 0.21$ and $\alpha = 0.04$, providing a mean absolute error of 4.24, an improvement over the previous

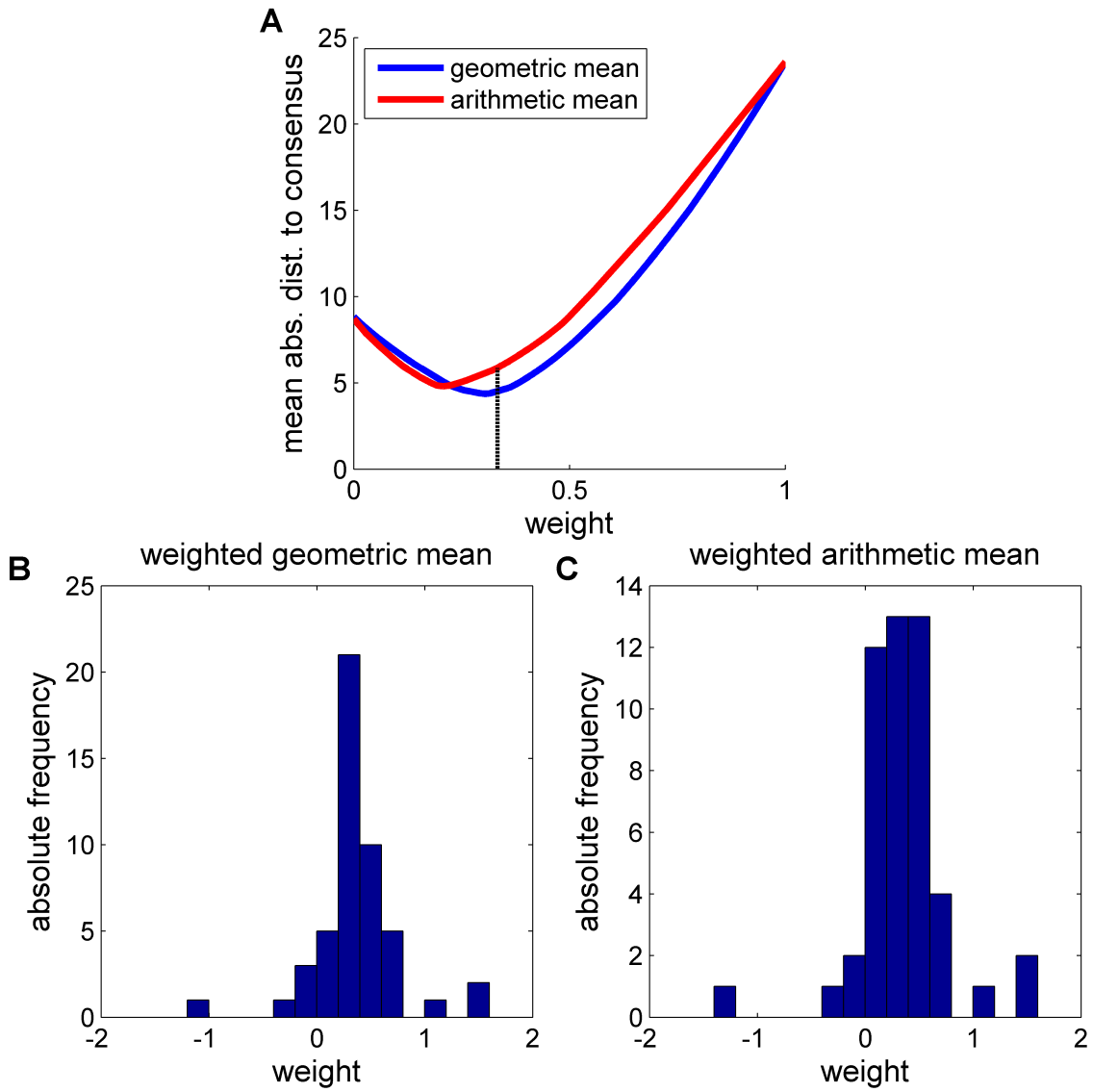

Fig. 3.2 **Consensus prediction with equal weights for the two closer estimates. A** Mean absolute difference between the real consensus value and the value predicted by Eqs 3.5 (geometric mean model, blue line) and Eqs 3.6 (arithmetic mean model, red line), with the same weight w applied for all groups. Vertical black line is $w = 1/3$. **B** Weights distribution in the geometric mean model, applying Eq 3.7 to each group. **C** Weights distribution in the arithmetic mean model, applying Eq 3.8 to each group.

methods. It can also be proposed a method like this, but based on the geometric mean. In this case, as the geometric mean is the exponential of the arithmetic mean of the logarithms,

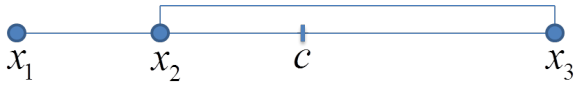

Fig. 3.3 **Spring model.** The effect of the outer estimate $x-1$ closer to the median one x_2 is like a spring force that pulls the consensus value c from the average of the median and the further estimate x_3.

it seems logic to compute the weights using the logarithm of the estimates:

$$c = x_2^{(1-w)} + x_3^w, \tag{3.11}$$

with the weight expressed as a function of two parameters and the distance between the intermediate estimate x_2 and the one closer to it x_1:

$$w = Ae^{-\alpha|\log(x_2)-\log(x_1)|}. \tag{3.12}$$

In this case, the best average prediction of consensus happens at the values of the parameters $A = 0.28$ and $\alpha = 0.93$, providing a mean absolute error of 4.10, an even better improvement than the weighted arithmetic mean.

3.2.4 Costs model

We finally tested a costs model, based on the concept that subjects reach a consensus that minimizes some cost function, that is, that minimizes the cost of changing from the individual pre-discussion estimates to a group common estimate. There are many different functional forms for a cost function applied successfully in optimization theory. Here we used a cost function that depends only on the distance from each subject initial guess to the final consensus and some strength parameter η. For three estimating subjects, the cost function can be expressed as:

$$cost = |c-x_1|^\eta + |c-x_2|^\eta + |c-x_3|^\eta \tag{3.13}$$

Given the three initial guesses, the optimal consensus (that is, the consensus value that minimizes the cost function) is a function only of the strength value. If we plot the distribution of the η value that best predicts the real consensus agreed by each group (Fig 3.4), two principal features are noticeable. One is that there exist some attractors, values of η that predict consensus for many groups. Other is that for many groups, no value can be found, which in the figure translates into $\eta = 5$, the maximum value tested, being an attractor.

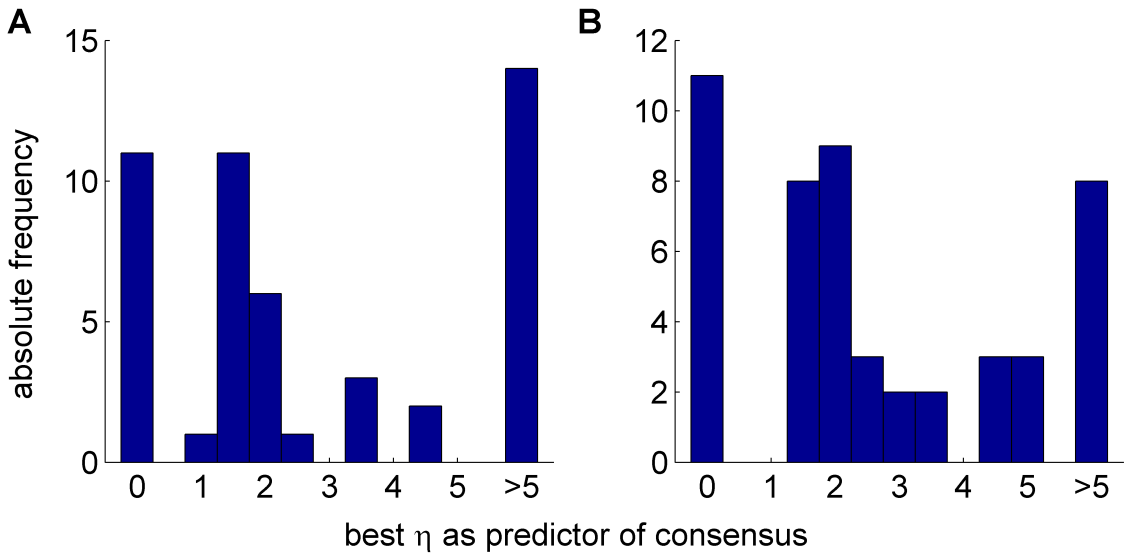

Fig. 3.4 **Distribution of the strength values η that best predict consensus for each group. A** Model using Eq 3.13. Peak of η between 1 and 2 explained in B. **B** Model using Eq 3.13, but with c and $x_{1,2,3}$ the logarithm of the estimates, which implies that the peak in $\eta = 2$ corresponds to the geometric mean of the estimates.

By inspection of Eq 3.13 it can be deduced that certain values of η correspond to special values of consensus c, precisely those corresponding to the attractors found: $\eta = 0$ corresponds to $c = x_2$ (or lower); $\eta = 2$ corresponds to $c = (x_1 + x_2 + x_3)/3$; and $\eta \to \infty$ to $c = (x_1 + x_3)/2$ (or higher).

3.2.5 Diversity in aggregation strategies

The results of the previous sections suggest that, despite the evidence that the geometric mean provides a good fit to the overall observed consensus estimates (Fig 3.5), it is feasible that different aggregation strategies were used to decide collectively, especially because the level of disagreement may influence how group decisions are made (Sniezek and Henry [29]).

We find that although the geometric rule appeared to be used most frequently, there is still room to consider the use of the other aggregation rules, with each alternative rule (Fig 3.6A) being the closest to the group consensus estimate at least twice. However, the frequency of using alternative rules, other than using the mean of the lowest and highest estimates, was within the 95% confidence intervals of assuming the geometric mean was used to aggregate initial estimates in each group, with a level of noise added that matched the noise in estimates in the experiment (Fig 3.6A).

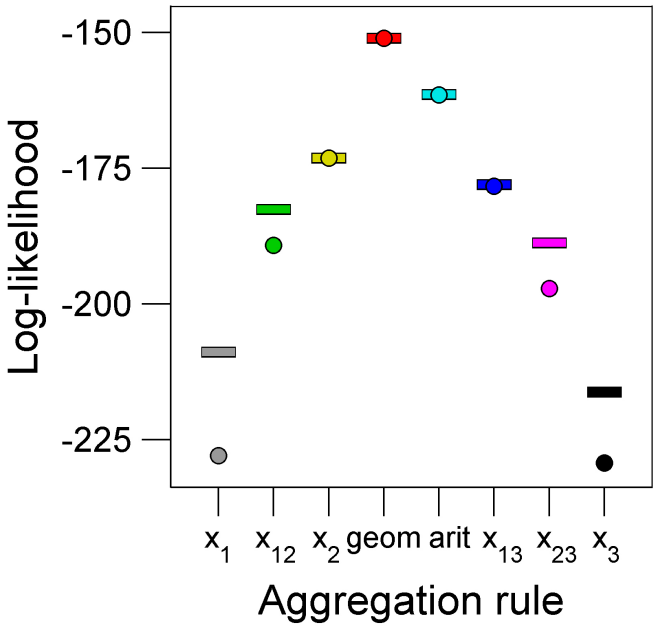

Fig. 3.5 **Fits of different aggregation rules to the observed group consensus estimates.** Observed log-likelihoods of eight different rules for aggregating initial estimates (circles) are plotted with the log-likelihoods when the noise (dashes) added to each estimation maximizes the log-likelihood. For the median (x_2, yellow), geometric (geom, red) and arithmetic (arit, cyan) means, and mean of the lowest and highest estimates in each group (x_{13}, blue), the fits to the data are close to maximal. The other rules tested are: the lowest estimate (x_1, gray), mean of the lowest and median estimate (x_{12}, green), mean of median and highest estimate (x_{23}, magenta), and highest estimate (x_3, black). The strategies are sorted in the x axis in an order that results in increasing values for many of the groups.

After splitting the data into groups with a low or high range of pre-discussion initial estimates (Fig 3.6B), there were a number of potential rules being applied in groups with low range, although it is more difficult to statistically distinguish between different rules when the range is smaller as, by definition, initial estimates are more similar to one another (See Fig 3.7B). At high ranges, the geometric mean was clearly the most common strategy (Fig 3.6C and Fig 3.7C).

We tested the consequences of using different aggregation rules for the accuracy of group decision making. For groups with low range, only the highest estimate and the average of the highest and median estimates significantly outperformed the geometric mean rule (Fig 3.6D). In contrast, for groups with high ranges, the geometric mean outperformed all alternatives. The rules that outperformed the geometric mean at low group ranges and the rule that was used more than expected compared to the noisy geometric rule at low ranges

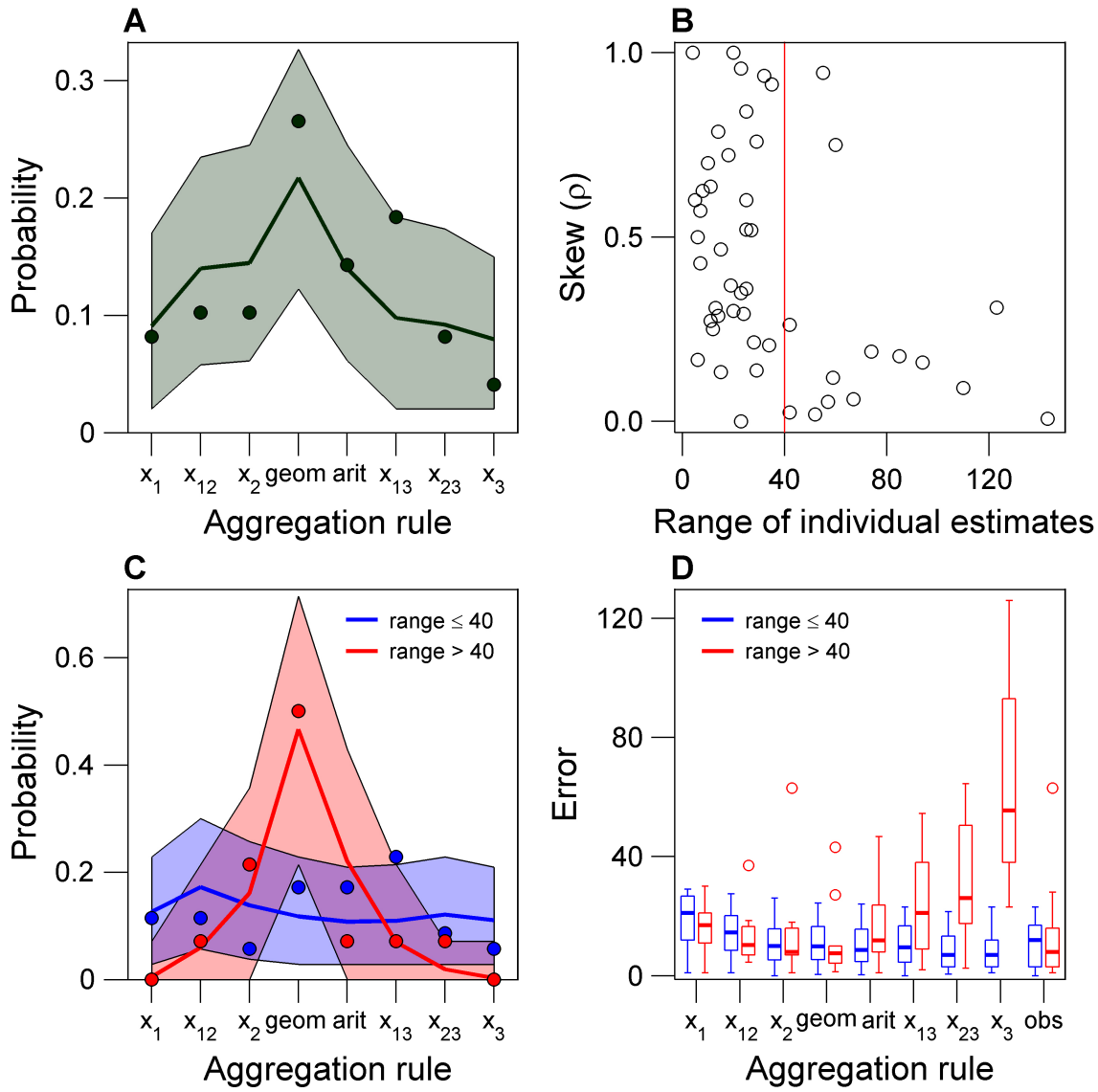

Fig. 3.6 **The use and consequences of different aggregation rules.** (**A**) Probability of each aggregation rule being the closest to the observed consensus estimates of the groups (filled circles). Also plotted is the probability (mean is black line, and shaded region is 95% confidence intervals) that the aggregation rule is the closest to the observed consensus estimates when a 'noisy' geometric mean simulation is instead used to aggregate the initial estimates (see Section 3.4.4). (**B**) Range and skew in initial estimates for each group. The range of initial estimates is plotted against the relative distribution of the estimates. Skew is $\rho \equiv (x_2 - x_1) / (x_3 - x_1)$, and is close to zero if the highest estimate (x_3) is a relative outlier, and close to one if lower estimate (x_1) is a relative outlier. The threshold between groups of low and high range (red line) is an approximate point that separates the region with any configuration (≤ 40) to the region with the two lower estimates being much closer to each other than to the higher (> 40). (**C**) As A, but separately for groups with a low range of estimates (blue dots, line and shaded area) and high range of estimates (red dots, line and shaded area). (**D**) Absolute error if each of the strategies had been followed exactly by groups with low range (blue) and high range (red). The notation for strategies as in Fig 3.5, with a final column added in D for the error of the observed group consensus estimates.

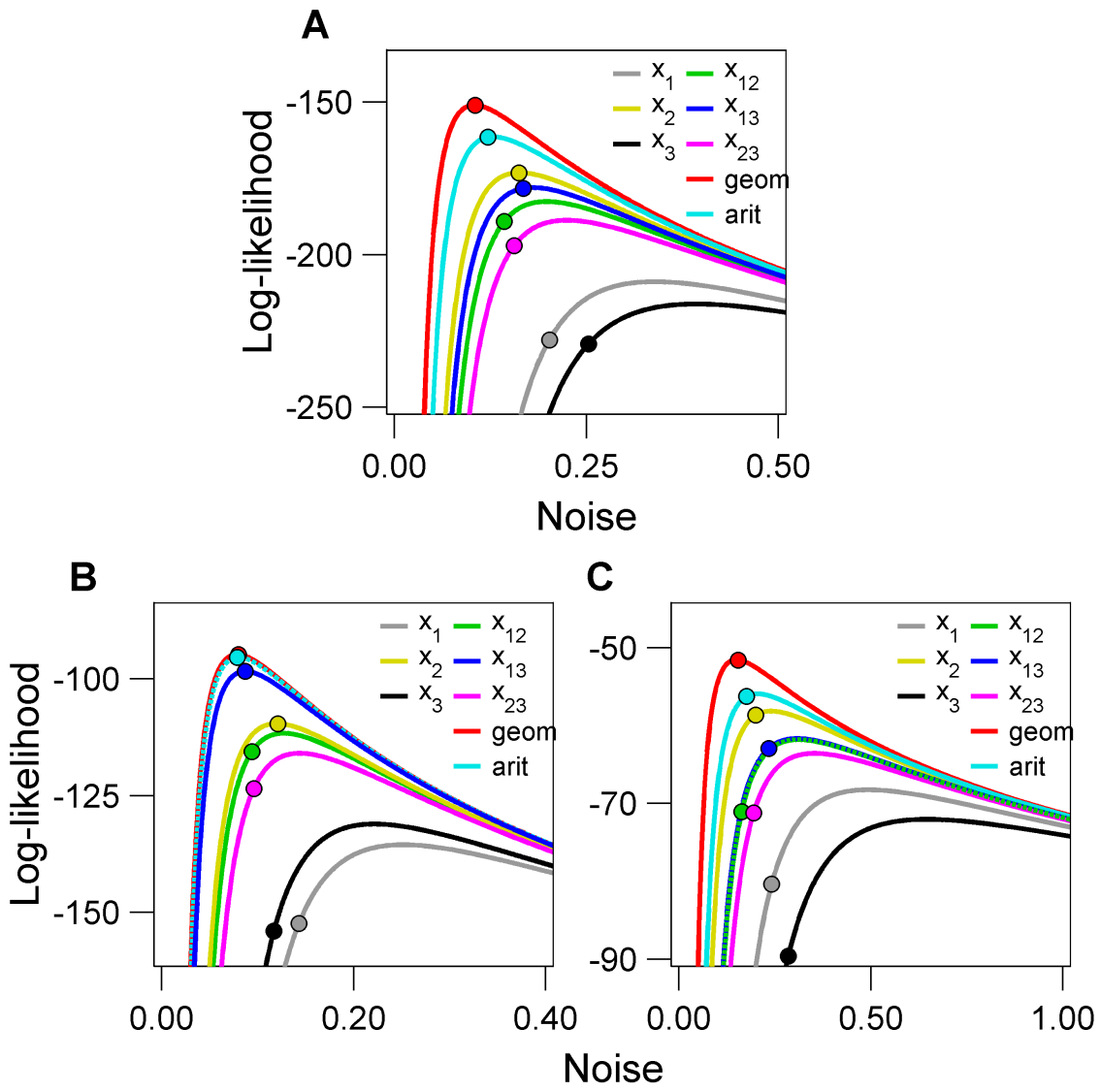

Fig. 3.7 **Fits of different aggregation rules to the observed data at various levels of added noise.** (A) Log-likelihood of the observed consensus values being produced from the exact value of each rule tested for a range of noise values (curves). Highlighted are the log-likelihood values corresponding to the observed level of deviation around the rule (circles). (B) As A, but only for groups with a low range (≤ 40) of estimates. (C) As A, but only for groups with a high range (> 40) of estimates. The rules tested are: the lowest estimate (x1), mean of the lowest and median estimate (x12), the median (x2), geometric (geom) and arithmetic (arit) means, mean of the lowest and highest estimates in each group (x13), mean of median and highest estimate (x23), and highest estimate (x3). At each level of noise, the geometric mean consistently provides amongst the best fits to the observed data.

gave particularly large errors in groups with high ranges (Fig 3.6D). Thus, the geometric mean provides a robust and generally high performing aggregation rule, particularly when there is disagreement in estimates within a group, and this trend matches its preferential usage (Fig 3.6C). There was no difference in accuracy between the geometric mean of the initial estimates and the group consensus estimates across all groups, or only in groups with low or high ranges of initial estimates.

The threshold used to define groups with low and high ranges did not have any effect on the trends in Fig 3.6C,D, that is, in the strategies distributions and accuracy of strategies (Fig 3.8).

3.3 Discussion

The prevalence of the geometric mean as the aggregation strategy suggests that the processing of information is done in a logarithmic-like fashion. Since the Weber-Fechner law was introduced, a logarithmic mental representation of numbers has been widely reported. This processing is more apparent in the data analyzed in the previous chapter, where the distributions of the estimates prior and post social interactions are log-normal. Here, although there are several high outlayers, the distribution of pre-discussion estimates is not significantly log-normal.

Although the noisy geometric mean model is the most likely strategy, there is clear stickiness to the exact values of strategies, which of course indicates that in a verbal discussion, all sort of strategies can arise, including adopting the opinion of one individual. The opinion of this individual can sometimes predominate because he is the more dominant, because it is the intermediate opinion ans the median looks more democratic for some, or because he convinces the others of his competence in the kind of tasks.

The experimental situation investigated in this chapter shows that the model presented in the previous chapter can be applied to situations in which subjects interact directly, and not only receive information about the opinions of the others. This shows that social interactions not only does not undermine collective wisdom, but enhances it. The experiment was performed with groups of three subjects, and it remains to be studied whether in higher group sizes the diversity in aggregation strategies is higher or lower.

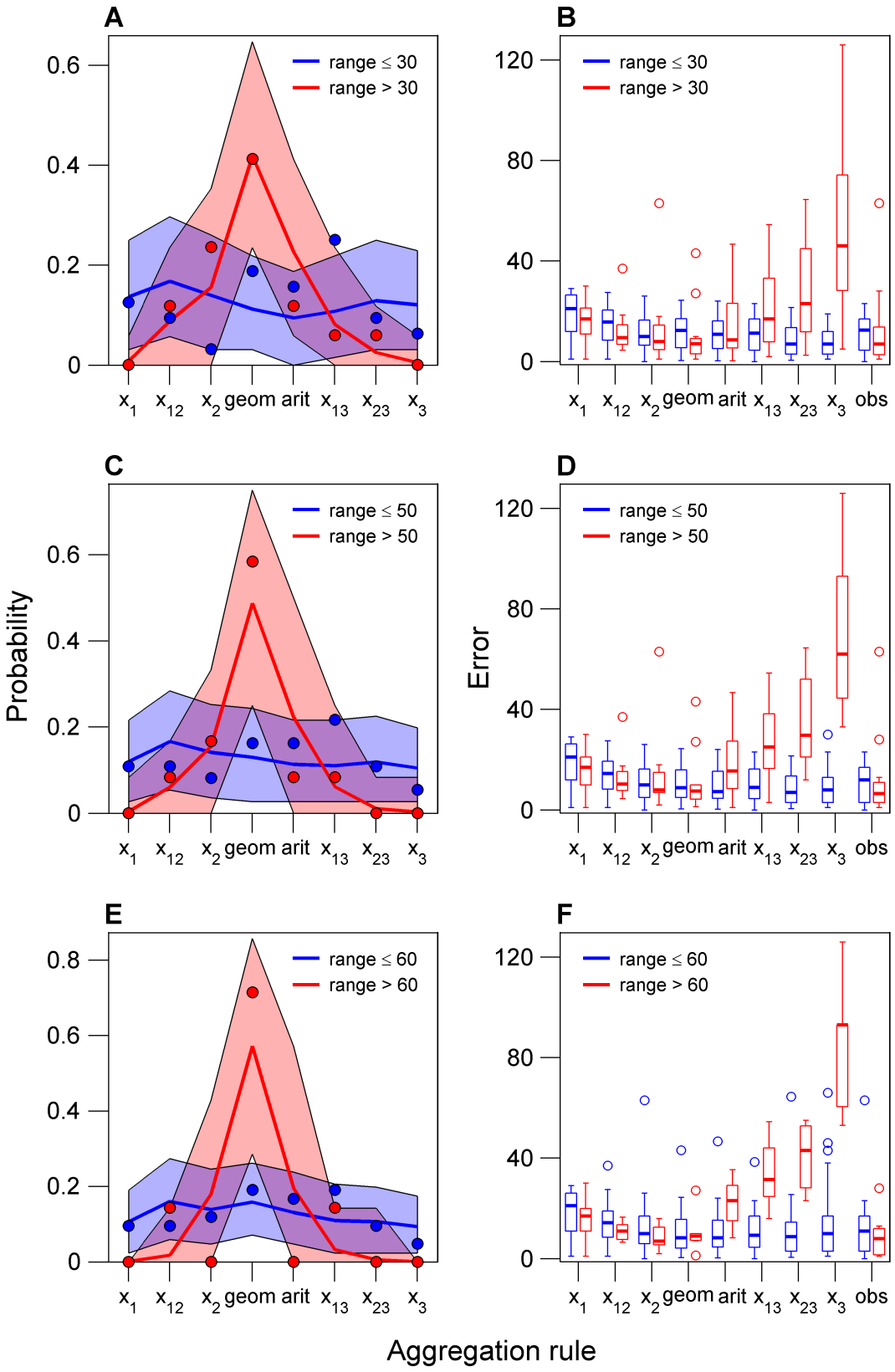

Fig. 3.8 **The use and consequence of different aggregation rules for different thresholds that define groups as having a low or high range.** The thresholds tested are 30 (**A**,**B**), 50 (**C**,**D**) and 60 (**E**,**F**). Plotting as in Fig 3.6C,D.

3.4 Materials and Methods

3.4.1 Data

The experimental situation consisted of 147 adolescent alumni from different schools and levels. They were asked to estimate the number of sweets in a crystal jar. The actual number of sweets was 57. They were first asked to estimate independently. Then they were arranged in groups of three, and were allowed to freely discuss. Each group was asked a consensus estimate. After that, subjects were asked to give a new independent estimation. We did not analyze the data from this las step.

3.4.2 Log-likelihood of simple aggregation rules

We computed the log-likelihood of each of the proposed aggregation rules as the logarithm of the likelihood L that the group consensus estimates are generated by a noisy computation of the rule. This was modeled by considering probability distributions centered at the values of the rule computed with the individual pre-discussion estimates of each group:

$$L\left(R|\left\{c_i\right\}_{i=1}^{n}\right)=f\left(\left\{c_i\right\}_{i=1}^{n}|R\right)=\prod_{i=1}^{n}f_i\left(c_i|r_i\right) \tag{3.14}$$

where R denotes one particular rule, $\{c_i\}_{i=1}^{n}$ is the set of observed discussion estimates agreed by each of the n groups, r_i is the exact value that the rule takes for the *i-th* group, and f_i was assumed to be a normal or log-normal with parameters r_i and σ. For each of the considered rules, we covered a wide range of possible values for σ to test for the dependence of the likelihood on the level of noise considered. The geometric mean was found to be the most likely rule to be generating the experimentally observed group estimates (Fig. 4), and the log-normal noise provided a higher log-likelihood value than the Gaussian noise (data not shown).

3.4.3 Noisy geometric mean model

We modeled groups of three subjects reaching a consensus from their initial individual estimates. Specifically, we considered that, given three estimates x_1, x_2 and x_3, the group

gave a consensus estimate c sampling from some probability density function

$$f\left(c|x_1,x_2,x_3\right). \tag{3.15}$$

In Section 3.4.2, we show that of all the rules proposed, the one with a higher likelihood of producing the experimental results is the geometric mean (Fig. 4):

$$f(c|g), \tag{3.16}$$

with

$$g \equiv \left(x_1 x_2 x_3\right)^{1/3} \tag{3.17}$$

the geometric mean of the three pre-discussion estimates. The noise that produces the consensus value to deviate from the geometric mean can have a Gaussian form,

$$f\left(c|g\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{c-g}{\sigma}\right)^2}, \tag{3.18}$$

or a log-normal form,

$$f\left(c|g\right) = \frac{1}{c\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(c)-\log(g)}{\sigma}\right)^2}. \tag{3.19}$$

We favored the log-normal option for two reasons. First, it provides a higher log-likelihood value (data not shown). Second, it is in greater agreement with the logarithmic-like effect shown in the main text and the nearly log-normal shape of the distribution (although not significant, Kolmogorov-Smirnov test over the z-scored logarithm of individual pre-discussion estimates, $p = 0.035$; also see Fig. S1D).

The σ parameter of the noise function can be established in at least two different ways. One is to use the value that provides a higher log-likelihood (Eq 3.14). The other is to compute the standard deviation of the set of 49 experimental values $\{\Delta_i\}_{i=1}^{49}$, where Δ_i is defined as the difference in logarithms from the geometric mean g_i of the *i-th* group's pre-discussion estimates to their group consensus estimate:

$$\Delta_i \equiv \log\left(c_i\right) - \log\left(g_i\right). \tag{3.20}$$

Both methods provide similar results (discrepancies of less than 1%).

The set of simulated consensus estimates $\{c_i\}_{i=1}^{49}$ will be thus a 49 dimension random variable, produced by the set of 49 probability density functions $\{f_i\}_{i=1}^{49}$, with

$$f_i\left(c_i|g_i\right) = \frac{1}{c_i\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(c_i)-\log(g_i)}{\sigma}\right)^2}, \tag{3.21}$$

for the case of log-normal noise. Note that uncorrelated groups are assumed.

3.4.4 Detailed protocol for creating Fig 3.6A,C using the noisy geometric mean model

We estimated the σ parameter to be used in Eq 3.21 computing the standard deviation of the set $\{\Delta_i\}_{i=1}^{49}$ of values obtained applying Eq 3.20 to the 49 experimental groups. With this standard deviation parameter, we generated a set $\{\xi_i\}_{i=1}^{49}$ of noise values to be added to each of the 49 geometric mean values g_i obtained from the pre-discussion estimates of the groups (Eq 3.17):

$$\log\left(c_i\right) = \log\left(g_i\right) + \xi_i \tag{3.22}$$

with each ξ_i value sampled from a Gaussian $N(0,\sigma)$. This way, we obtained a set $\{c_i\}_{i=1}^{49}$ of 49 simulated consensus values, and then determined for each group which rule (Fig. 4) was closest to their simulated group consensus estimate. For the *i-th* group, the set $\left\{r_{ij}\right\}_{j=1}^{8}$ of values that each of the 8 considered rules take was computed from the pre-consensus values obtained in the real experiment. To assign one or more rules from the set $\left\{r_{ij}\right\}_{j=1}^{8}$, a minimum distance criterion was applied. For that, the set $\left\{d_{ij}\right\}_{j=1}^{8}$ of distances from the generated consensus value to that of the rule,

$$d_{ij} \equiv \left|c_i - r_{ij}\right|, \tag{3.23}$$

was computed. Then, the *j-th* rule was classified as followed by the *i-th* group if

$$d_{ij} = \min\left\{d_{ij}\right\}_{j=1}^{8}. \tag{3.24}$$

The frequency q_j of the *j-th* rule was not computed simply as the number of groups for which Eq 3.24 was fulfilled by d_{ij}, because for some groups Eq 3.24 was fulfilled by not one

but n_i rules. Instead, the contribution of the *i-th* group to each of the rules was determined as

$$q_{ij} = \begin{cases} 1/n_i & \text{if } d_{ij} \text{ fulfills Eq 3.24} \\ 0 & \text{otherwise} \end{cases}, \tag{3.25}$$

and then the frequencies were actually computed as

$$q_j = \sum_{i=1}^{49} q_{ij}. \tag{3.26}$$

Note that, for each group, we did not sort the rules in ascending numerical order, but keeping always the same operation performed with the three pre-consensus values in the same position across groups. To turn the frequencies q_j into probabilities p_j, we divided each by the total number of rules considered:

$$p_i = \frac{q_i}{8}. \tag{3.27}$$

We repeated 10,000 times the process we have detailed, obtaining for each rule a sample distribution of 10,000 frequencies (and probabilities computed with Eq 3.27 when required) compatible with the noisy geometric model. For each of these distributions, the mean and 2.5 and 97.5 percentiles were computed. This way, we obtained for each rule the mean of compatible probabilities (blue line in Fig 5A,C), and the limits that contain 95% of compatible probabilities (upper and lower limits of the shaded areas in Fig 5A,C).

Chapter 4

Neural Networks to Improve Diagnosis in Groups of Doctors

4.1 Introduction

In the previous chapters we have applied aggregation techniques to find improvements over straightforward averaging of estimates, or studied under which circumstances a discussing group is able to detect when averaging is the best strategy given the individual pre-discussion estimates. In the absence of more information, and assuming errors normally distributed, the arithmetic mean is the aggregate that minimizes the expected error of collective estimates. In the case of log-normal distribution of errors, the geometric mean is the optimal strategy.

Although the usual situation in everyday life is one single doctor facing a clinical case, the approach of the wisdom of a collective is very suitable to disease diagnosis, and is been so applied in Kurvers et al. [14]. Other selection strategies (like the confidence rule, or selection by accuracy in previous cases), and even a combination of them have been applied to the classification by humans problems. However, there might be subtleties in the assessment of the quality of each piece of information that a human, or a simple heuristic applied by him like majority voting is not able to capture.

For example, correlations in decisions by individuals are not taken into account in any of the aforementioned strategies, and might be essential when setting decision boundaries. For example, correlations have been used successfully in the modeling and prediction of information processing by big groups of neurons.

Neural networks have been shown to approximate any function, given a sufficiently complex structure of layers. Because of that, they are being applied to an increasingly high number of classification problems. In fact, the applications of neural networks are going beyond simple classification, and are being applied to complex tasks like autonomous cars.

4.2 Results

4.2.1 Comparison of Network and Heuristics in Pairs of Doctors

We examined whether neural networks could improve upon the performance of previously proposed heuristics on a skin cancer classification dataset (Kurvers et al. [14]). In the dataset, explained in more detail in Section 4.4.1, 40 doctors had given their estimations and confidence scores on whether 108 patients had malignant melanoma by examining images of their skin lesions. As in Kurvers et al. [14], we used Youden's index J as a measure of accuracy, given by J=sensitivity + specificity - 1, with sensitivity defined as the proportion of positive cases correctly and specificity defined as the proportion of negative cases correctly evaluated (see Section 4.4.5). This measure weights equally sensitivity and specificity and it is thus insensitive to the unbalances of a dataset (in this case, more cases without cancer than with cancer). We then generated virtual pairs of doctors and examined the accuracy of their aggregated judgments. If the two doctors agreed on a diagnosis, their joint shared opinion was used as the diagnosis. If there was disagreement, we compared the performance of the following three heuristics for conflict resolution:

1. Use the opinion of the more accurate doctor in the pair ('best').
2. Use the opinion of the more confident doctor ('confident').
3. If the accuracy of both doctors was similar ($|\Delta J| < 0.1$), we used the 'confident' rule. If accuracies were not similar, we used the 'best' rule ('conf/best').

The three described heuristics are applied in Kurvers et al. [14] to groups of two, three and five doctors, and is there shown that 'conf/best' rule heuristic outperforms the other two heuristics for each group size.

We also fitted a neural network that was given as input the historical accuracies of the two doctors and their declared confidence scores (see Section 4.4.2). We asked whether a network can find an aggregation decision rule better than the three proposed heuristics. The

network was trained with 50% of the skin lesion cases and validated with the remaining 50% (see Fig 4.1A for the training of one such network).

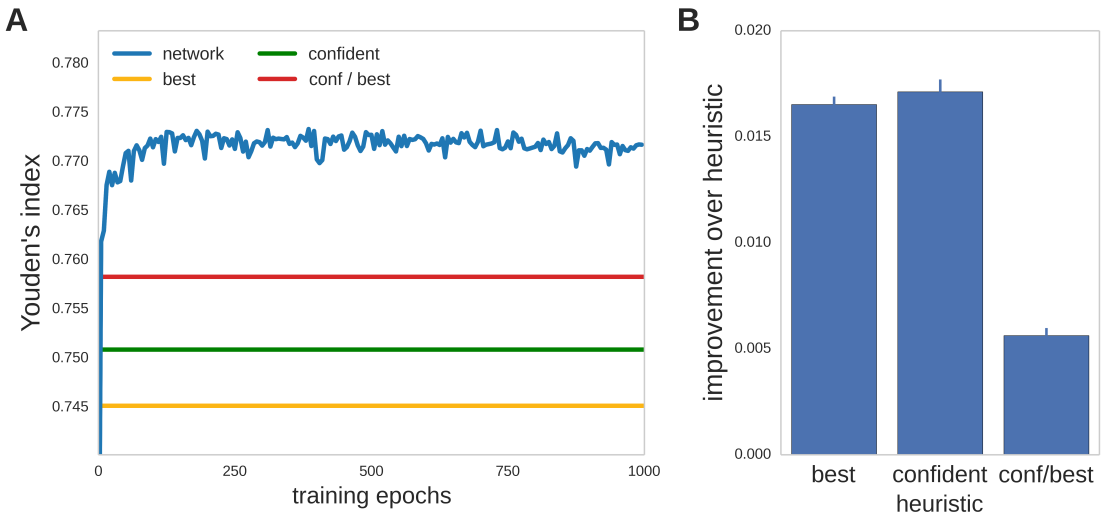

Fig. 4.1 **Network learning how to combine the opinions and confidence scores of two doctors into a cancer/no cancer classification rule.** (**A**) Example of the training of one network for a particular partition of the data into 50% for training and 50% for validation. Shown is the evolution of the performance for the validation data (Youden's index, J =sensitivity + specificity -1) of the network (blue line), and the performance of the best doctor heuristic (yellow line), the more confident heuristic (green line), and the conf/best heuristic (red line). (**B**) Mean improvement in Youden's index of the network over the heuristics. Error bars are sem.

We trained 300 networks using different $50-50\%$ partitions of the data into training and validation. We found mean network performance of $J = 0.740$ and standard deviation of 0.036. The different heuristics had the following performance for the same data: 0.723 ± 0.034 ('best'), 0.722 ± 0.038 ('confident') and 0.734 ± 0.035 ('conf/best'); see Fig 4.1B for mean improvement of network over heuristics. We found that all the 300 networks were better than the three heuristics, resulting in $p < 10^{-7}$ for the 'best' and 'confident' heuristics using a Wilcoxon rank-sum test, and $p = 0.046$ for the 'conf/best' heuristic with the same test.

4.2.2 Performance and group size

We then investigated whether the network consistently outperformed over simple heuristics for higher group sizes. To do that, we created virtual groups of three, five and seven doctors,

and computed their performance with the heuristics proposed in Section 4.2.1. We included two new heuristics:

1. Use the opinion held by a higher number of doctors ('majority').

2. If the accuracy of both doctors was similar ($|\Delta J| < 0.1$), we used the 'majority' rule. If accuracies were not similar, we used the 'best' rule ('maj/best').

These heuristics are also proposed in Kurvers et al. [14] for groups of three and five doctors, and is there shown that the 'maj/best' heuristic outperforms the 'majority' and the 'best' heuristics for each group size.

Like in Section 4.2.1, we fitted a neural networks that were given as input the historical accuracies of each doctor of the groups and their declared confidence scores (see Section 4.4.2). The cases were again divided in 50% of them for training and the remaining 50% for validation. For each group size, we trained 250 different networks using different $50-50\%$ partitions of the data into training and validation.

We found that both the network and the five heuristics proposed improved their performance over the validation cases for increasing group sizes (Fig 4.2A). We also found that all the 250 networks were better than the five heuristics for all the group sizes analyzed (Fig 4.2B)

4.3 Discussion

The strategy that combines the confidence criterion with the best doctor criterion performs worse than the confidence strategy for group sizes of 3 and 5. Also, the strategy that combines the majority criterion with the best doctor criterion preforms worse than the majority voting criterion for all the group sizes tested. On the contrary, in Kurvers et al. [14] the combining strategies are reported to beat the simple ones for all the group sizes tested there (2, 3 and 5). The explanation for this mismatch is that whereas in Kurvers et al. [14] the accuracy of each doctor is computed over all the experimental cases, we computed it only over the training cases. To guarantee the predictive power of a model, it is essential not using any numerical result that includes data from the testing set, for making prediction over the same validation cases.

It is in the spirit of many collective intelligence studies, and also in accordance with the results shown in Chapter 2 that the majority voting strategy is the only one that has a close

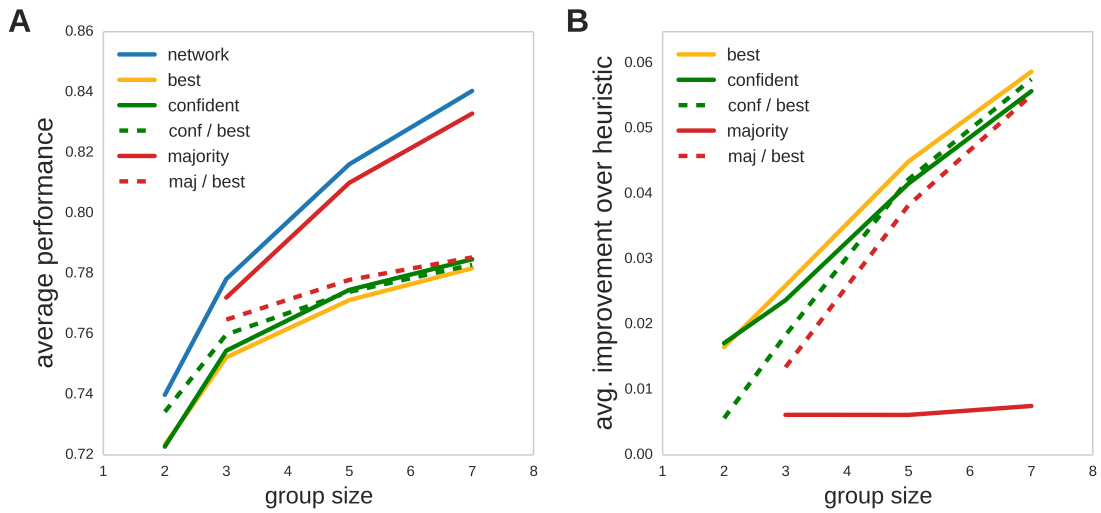

Fig. 4.2 **Comparison of performances of neural network and aggregation heuristics for different group sizes.** (**A**) Average of the performance of network (blue line) and heuristics over all the validation sets, for each of the group sizes.(**B**) Average improvement of the network over the heuristics in the validation sets, for each of the group sizes. The strategies (or their distance to the network) are choosing the best doctor (yellow), the more confident one (green), the majority voting (red), the combined confident and best (dashed green) and the combined majority and best (dashed red). The group sizes tested were 2, 3, 5 and 7.

performance to the network. This suggests that, although the network finds more complex structure in the data than that implied in the formulation of simpler aggregation rules, a majority voting is in the basis of its computation.

Some deviations from the origin of the data fed to the network could be introduced for further developments of the method proposed here. It would be interesting to use as data not (or not only) the accuracy of the estimators in an almost identical set of data, but maybe in less closely related tasks, or subjective ratings like the level of satisfaction in the hospital where they worked or by patients they treated. Also, use non experts instead of professionals, and test whether there is a higher predominance of the 'best doctor' strategy or the majority voting is still the basis of the network learning.

4.4 Matherials and Methods

4.4.1 Data

We used the data analyzed in Kurvers et al. [14], that can be downloaded in http://www.pnas.org/content/113/31/8777.full?tab=ds. This dataset is comprised of evaluations of 40 doctors on 108 different cases of potential melanomas. For each case, the doctors were also asked to declare confidence in their personal judgment in a 1 to 4 scale.

4.4.2 Input data

We performed a different partition into 54 training and 54 validation cases each time we trained a network. We trained 300 different networks for a group size of 2 doctors, and 250 networks for group sizes of 3, 5 and 7 doctors. To train each network, we generated random virtual groups up to a maximum of 1000 groups, in a similar manner as in Kurvers et al. [14]. The total number of combinations of 40 doctors extracted in groups of 2 is 780, so for group size of 2 all possible combinations were selected. For higher group sizes the number of combinations is higher than 1000, so the groups used for the training and computing the average performance of the heuristics were selected at random.

Accuracy of each doctor was determined computing his Youden's index (J=sensitivity + specificity -1) over the training cases. We computed the accuracy of each group using the different heuristics proposed over the validation cases. The performance of the heuristics was then determined by averaging its value across all groups.

To train each network, we generated training instances combining judgments, accuracies and confidence ratings of the doctors of the group on each particular case. For example, for a group size of three doctors, each input was then composed of accuracy of first doctor, confidence of first doctor, accuracy of second doctor, confidence of second doctor, accuracy of third doctor, and confidence of third doctor. Accuracies were multiplied by -1 if the doctor had judged the case as negative. We used the training cases to train the network and the validation cases to compare its performance with the heuristics applied to the groups. The training and validation data sets for the example of groups of three doctors is composed

of $1000 \times 54 = 54000$ six-tuples of the kind:

$$x = \begin{pmatrix} x_1^1 \\ x_2^1 \\ \vdots \\ x_6^1 \end{pmatrix} = \begin{pmatrix} \text{judgment 1} \times \text{accuracy 1} \\ \text{confidence 1} \\ \text{judgment 2} \times \text{accuracy 2} \\ \text{confidence 2} \\ \text{judgment 3} \times \text{accuracy 3} \\ \text{confidence 3} \end{pmatrix}. \tag{4.1}$$

4.4.3 Neural network architecture

Once the input data is constructed, it is introduced in the neural network algorithm. Fig 4.3 shows a neural network with an input example of N elements (with $N = 2 \times$group size) like described in Eq 4.1 feeding the input layer, with each element of the example assigned to the numerical value of one input neuron (Fig 4.3, x values). The input layer is therefore composed of N neurons. Then the value of each input neuron is sent to each of the M neurons

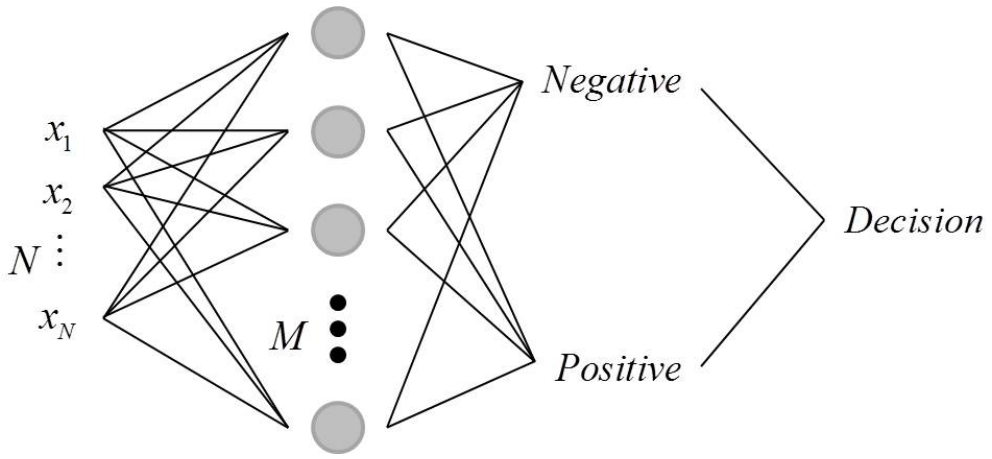

Fig. 4.3 **Diagram of a neural network.** The network receives data of size N, has a hidden layer composed of M neurons, and a binary decision layer.

of the hidden layer (Fig 4.3, gray circles). The value assigned to each neuron is then a linear combination of the value of the neurons of the previous layer:

$$\begin{pmatrix} w_{11}^1 & w_{12}^1 & \cdots & w_{1N}^1 \\ w_{21}^1 & w_{22}^1 & \cdots & w_{2N}^1 \\ \vdots & \vdots & \ddots & \vdots \\ w_{M1}^1 & w_{M2}^1 & \cdots & w_{MN}^1 \end{pmatrix} \begin{pmatrix} x_1^1 \\ x_2^1 \\ \vdots \\ x_N^1 \end{pmatrix} + \begin{pmatrix} b_1^1 \\ b_2^1 \\ \vdots \\ b_M^1 \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^N \left(w_{1n}^1 x_n^1\right) + b_1^1 \\ \sum_{n=1}^N \left(w_{2n}^1 x_n^1\right) + b_2^1 \\ \vdots \\ \sum_{n=1}^N \left(w_{Mn}^1 x_n^1\right) + b_M^1 \end{pmatrix} \xrightarrow{\text{ReLU6}} \begin{pmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_M^2 \end{pmatrix} \tag{4.2}$$

with the final ReLU6 representing a rectified linear unit:

$$\text{relu}\,(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 6 \\ 6 & \text{if } x > 6 \end{cases} \tag{4.3}$$

The same process can repeated with the appropriate number of hidden layers. The values of the neurons of the last hidden layer are sent to the output layer (Fig 4.3, *negative*, *positive*):

$$\begin{pmatrix} w_{11}^2 & w_{12}^2 & \cdots & w_{1M}^2 \\ w_{21}^2 & w_{22}^2 & \cdots & w_{2M}^2 \end{pmatrix} \begin{pmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_M^2 \end{pmatrix} + \begin{pmatrix} b_1^2 \\ b_2^2 \end{pmatrix} = \begin{pmatrix} \sum_{m=1}^M \left(w_{1m}^2 x_m^2\right) + b_1^2 \\ \sum_{m=1}^M \left(w_{2m}^2 x_m^2\right) + b_2^2 \end{pmatrix} \xrightarrow{\text{softmax}} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}. \tag{4.4}$$

Instead of applying a rectified linear unit to the two linear combinations of the last layer, the softmax function of them is computed:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{i=1}^n e^{x_i}} \tag{4.5}$$

This way, two values that sum up to 1 are obtained. The decision about the example introduced in the input layer is then the one to which the higher value belongs.

In our case, for every group size the architecture of the networks consisted of two hidden layers, both of size 100 with rectified linear activation.

4.4.4 Network learning

The network was trained with the ADAM algorithm (Kingma and Ba [10]). The learning rates used were 0.0001 for the case of 2 doctors, 0.00001 for the case of 3 doctors, and 0.000001 for the cases of 5 and 7 doctors.

4.4.5 Loss function

The cost function was selected to match the accuracy measured by the Youden's index, which is defined as

$$J = \mathrm{TP}/(\mathrm{TP+FN}) + \mathrm{TN}/(\mathrm{TN+FP}) - 1, \tag{4.6}$$

with TP standing for true positives, FN for false negatives, TN for true negatives and FP for false positives. As the output of the network was the probabilities p and $1-p$ depending that the case fed was a positive or a negative, the expected value Y of the Youden's index would be of the form

$$Y \equiv E[J] = \frac{1}{n_p}\sum_{i=1}^{n_p} p_i + \frac{1}{n_n}\sum_{i=n_p+1}^{n_p+n_n}(1-p_i) - 1, \tag{4.7}$$

where n_p (n_n) is the number of positives (negatives), and the first (second) sum is over the positive (negative) cases. Like the Youden's index, the expected value lies in the interval $[-1,1]$, and therefore a loss function defined as

$$L \equiv \frac{1}{2}(1-Y) \tag{4.8}$$

The loss function is 0 at the maximum expected Youden's index ($E[J]=1$) and 1 at the minimum Youden's index ($E[J]=-1$).

Although Y is a natural adaptation of the Youden's index J to probabilities, the match between both is not perfect. This is caused by the fact that Youden's index is not continuous, based on dichotomic decisions, while the function Y is based in continuous probabilities. This gives rise to situations in which a better classification score gives however poorer probabilities (see Table 4.1 for an illustration).

Truth	p_i	Decision
P	0.6	P
P	0.6	P
N	0.4	N
N	0.4	N
N	0.4	N
	$Y : \sum p_i/n_p + \sum(1-p_i)/n_n - 1 = 0.2$	$J : \mathrm{TP}/n_p + \mathrm{TN}/n_n - 1 = 1$
P	0.9	P
P	0.9	P
N	0.1	N
N	0.1	N
N	0.9	P
	$Y : \sum p_i/n_p + \sum(1-p_i)/n_n - 1 = 0.53$	$J : \mathrm{TP}/n_p + \mathrm{TN}/n_n - 1 = 0.67$

Table 4.1 **Loss problem.** Illustration of the mismatch between Youden's index (J) and its predictor (Y). The correct classification is given in the 'Truth' column (P: positive, N: negative). The probability of the case being positive computed by the network is given in the column 'p_i'. According to the probability computed, a decision is taken in column 'Decision'. Both Y and J are computed separately for the first and last five cases. The disagreement is that with the five first cases the classification is perfect and in the second five it is not, but the loss function is worse with the first than with the last.

Chapter 5

Conclusions

- Bayesian estimation and probability matching provide an appropriate framework to model the integration of information received from a collective with the private (prior) information already possessed by the subject. The model predicts a weighted geometric mean of private and social information, which is found to be in good accordance with the observed change in the distribution of estimates under social influence.

- Although the model fits the data well at collective level, individuality is apparent within groups of subjects. This individuality in resistance to social influence is not homogeneously distributed, but often clusters of opinions are found in a group of independent subjects.

- Those individuals that rely more on their previous initial opinions are found to be on average more accurate than the full crowd, and clusters of resisting individuals are found often closer to the true value, and never further. Declared confidence is not a good predictor of competence, and is not found to be correlated with resistance to social influence.

- In a group discussion experiment, the geometric mean is found to be the more likely aggregation strategy followed by groups of three. This is in accordance with the prediction of the Bayesian model, and with the fact that in the kind of experiments analyzed the distributions of independent estimates are approximately log-normal.

- Although the geometric mean is the most likely aggregation rule if noise is included, there is diversity among groups in the strategy used to integrate the pre-discussion

opinions. Particularly, the geometric mean is extremely preferred when there is a higher outlayer estimate in the group.

- When distribution of estimates is symmetric and the range is low, all the central aggregation strategies provide similar accuracies. On the other hand, when there is asymmetry with a high outlayers, the geometric mean is the more accurate strategy, showing that groups integrate their individual opinions in way that improves their collective intelligence.

- Machine Learning tools like Neural Networks can be applied to integrate the opinions of a group of experts. The network is able to aggregate the estimates of the group to provide higher classification accuracy than standard decision strategies, like majority voting or choosing the opinion of the doctor that declares more confidence in his diagnosis.

- Although the performance of the standard aggregation heuristics increases with group size, the performance of the network grows faster than all of them, except the majority voting criterion.

- The results give open both the possibility of applying novel machine learning techniques to help improve over the decisions of experts, and the possibility to find refined aggregation of opinions rule that take more into account the correlations between individual estimators.

Chapter 6

Conclusiones

- La estimación Bayesiana y 'probability matching' proporcionan un marco adecuado para modelar la integración de la información recibida desde un colectivo con la información privada que el sujeto ya poseía. El modelo predice una media geométrica ponderada de la información privada y social, lo que está en buen acuerdo con el cambio que se observa experimentalmente en la distribución de estimaciones bajo influencia social.

- A pesar de que el modelo ajusta bien con los datos experimentales a nivel del colectivo, es notable la presencia de individualidad dentro del grupo. Esta individualidad en la resistencia a la información social no está homogéneamente distribuida, sino que a menudo se producen 'clusters' de opinión en grupos de sujetos independientes.

- Aquellos individuos que confían más en su opinión inicial previa son en promedio más precisos que el grupo entero, y los clusters de individuos resistentes están a menudo más cerca del valor correcto, y nunca más lejos. La confianza declarada no es un buen predictor de la precisión, y no se encuentra correlacionada con la resistencia a la influencia social.

- En un experimento de discusión en grupo, se encuentra que la media geométrica es la estrategia seguida con mayor plausibilidad en grupos de tres sujetos. Eso está en acuerdo con la predicción del modelo Bayesiano, y con el hecho de que en el tipo de experimentos analizados las distribuciones de estimaciones independientes son log-normales.

- A pesar de que la media geométrica es la regla de agregación más plausible si se tiene en cuenta el ruido, hay una diversidad entre grupos en la estrategia utilizada para integrar las opiniones pre-discusión. En particular, la media geométrica es claramente la preferida cuando hay una estimación extrema más alta en el grupo.

- Cuando la distribución de las estimaciones es simétrica y el rango es pequeño, todas las estrategias de agregación centrales proporcionan precisiones similares. Sin embargo, cuando hay asimetría y valores extremos altos, la media geométrica es la estrategia más precisa, mostrando que los grupos integran las estimaciones individuales de un modo que mejora la inteligencia colectiva.

- Herramientas de 'Machine Learning' como las Redes Neuronales pueden ser aplicadas para integrar las opiniones de un grupo de expertos. La red es capaz de integrar las estimaciones del grupo de un modo que proporciona una mayor precisión en la clasificación que la que dan estrategias de decisión habituales, como elegir la opinión mayoritaria o la del médico que declara una mayor confianza en su diagnóstico.

- A pesar de que la precisión de las reglas de agregación habituales aumenta con el tamaño del grupo, la precisión de la red aumenta más rápido que todas ellas, excepto que el criterio de la opinión mayoritaria.

- Los resultados abren tanto la posibilidad de aplicar nuevas técnicas de machine learning para ayudar a mejorar a partir de la opinión de expertos, como la posibilidad de encontrar reglas de agregación refinadas, que tengan más en cuenta las correlaciones entre los estimadores individuales.

References

[1] Arganda, S., Perez-Escudero, A., and de Polavieja, G. G. (2012). A common rule for decision making in animal collectives across species. *Proceedings of the National Academy of Sciences*, 109(50):20508–20513.

[2] Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., and Frith, C. D. (2010). Optimally Interacting Minds. *Science*, 329(5995):1081–1085.

[3] Bilmes, J. A. and others (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4(510):126.

[4] Bromiley, P. (2003). Products and convolutions of gaussian probability density functions. *Tina-Vision Memo*, 3(4).

[5] Budescu, D. V. and Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280.

[6] Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

[7] Galton, F. (1907). Vox populi. *Nature*, 75(7):450–451.

[8] Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Probability Distributions,(Vol. 1)*. John Wiley & Sons Inc., NY.

[9] King, A. J., Cheng, L., Starke, S. D., and Myatt, J. P. (2012). Is the true 'wisdom of the crowd' to copy successful individuals? *Biology Letters*, 8(2):197–200.

[10] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[11] Koriat, A. (2012). When Are Two Heads Better than One and Why? *Science*, 336(6079):360–362.

[12] Krause, J., Ruxton, G. D., and Krause, S. (2010). Swarm intelligence in animals and humans. *Trends in Ecology & Evolution*, 25(1):28–34.

[13] Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573–603.

[14] Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., and Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, page 201601827.

[15] Lee, M. D. and Shi, J. (2010). The accuracy of small-group estimation and the wisdom of crowds. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1124–1129.

[16] Lee, M. D., Steyvers, M., de Young, M., and Miller, B. (2012). Inferring Expertise in Knowledge and Prediction Ranking Tasks: Topics in Cognitive Science. *Topics in Cognitive Science*, 4(1):151–163.

[17] Limpert, E., Stahel, W. A., and Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question. *BioScience*, 51(5):341–352.

[18] Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.

[19] Mahmoodi, A., Bang, D., Ahmadabadi, M. N., and Bahrami, B. (2013). Learning to Make Collective Decisions: The Impact of Confidence Escalation. *PLoS ONE*, 8(12):e81195.

[20] Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276.

[21] Mavrodiev, P., Tessone, C. J., and Schweitzer, F. (2013). Quantifying the effects of social influence. *Scientific Reports*, 3:1360.

[22] McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

[23] Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press.

[24] Papoulis, A. and Pillai, S. U. (2002). *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education.

[25] Parkin, T. B. and Robinson, J. A. (1993). Statistical evaluation of median estimators for lognormally distributed variables. *Soil Science Society of America Journal*, 57(2):317–323.

[26] Pérez-Escudero, A. and de Polavieja, G. G. (2011). Collective Animal Behavior from Bayesian Estimation and Probability Matching. *PLoS Computational Biology*, 7(11):e1002282.

[27] Ribeiro, M. I. (2004). Gaussian probability density functions: Properties and error characterization. *Institute for Systems and Robotics, Lisboa, Portugal*.

[28] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

[29] Sniezek, J. A. and Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational behavior and human decision processes*, 43(1):1–28.

[30] Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

[31] Vul, E. and Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7):645–647.

[32] Wagner, C. and Vinaimont, T. (2010). Evaluating the wisdom of crowds. *Proceedings of Issues in Information Systems*, 11(1):724–732.

[33] Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R., and Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043.

[34] Wolfers, J. and Zitzewitz, E. (2004). Prediction markets. *The Journal of Economic Perspectives*, 18(2):107–126.

[35] Zhou, D., Basu, S., Mao, Y., and Platt, J. C. (2012). Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, pages 2195–2203.